Projet d'Analyse de Données Télémétriques pour l'Assurance (Data cleaning + EDA + Econométrie)

Introduction

Dans un monde où les assureurs sont constamment confrontés à des défis liés à la précision des prédictions et à l'optimisation des primes, la capacité à analyser et exploiter efficacement les données devient un atout majeur. Ce rapport vise à explorer les nouvelles bases de données d'assurance à travers une approche rigoureuse de data science. Nous avons pour objectif de comprendre les facteurs déterminants du montant des réclamations et de proposer un modèle prédictif fiable pour la tarification.

En utilisant uniquement Python, nous avons travaillé sur la préparation, l'analyse et la modélisation des données pour obtenir une compréhension approfondie des comportements des assurés. Les objectifs principaux de cette étude sont :

- Prédiction des Réclamations : Utiliser des techniques de régression et de machine learning pour anticiper les montants futurs des réclamations.
- Classification des Assurés : Catégoriser les réclamations en "élevées" ou "faibles" pour identifier les clients à risque.
- Compréhension des Variables Clés : Dégager les variables qui influencent le plus le montant des réclamations.

Ce rapport présente en détail l'ensemble du processus suivi, depuis la compréhension des données et leur préparation, jusqu'à la modélisation et l'évaluation des résultats. En analysant et en interprétant les résultats, nous proposons également des recommandations concrètes pour améliorer la précision des modèles et renforcer les stratégies de tarification.

L'approche que nous avons adoptée est divisée en étapes distinctes :

- Préparation et Traitement des Données : Compréhension de la qualité des données et traitement des anomalies.
- Analyses Descriptives et Graphiques : Exploration univariée et multivariée pour dégager les tendances.
- Modélisation : Modèles de régression et de classification pour répondre aux objectifs prédéfinis.
- Analyse des Résultats : Explication des résultats obtenus et identification des améliorations futures.

Ce rapport constitue un guide pour comprendre les principaux facteurs affectant les réclamations d'assurance et pour développer des modèles plus précis et plus robustes.

A- DESCRIPTION DES DONNEES

La période d'observation est comprise entre 2013 et 2016.

1. claims_data (DB_SIN.txt)

Cette base de données contient des informations détaillées sur les réclamations faites par les assurés. Les variables principales que l'on peut s'attendre à trouver incluent :

- **Id_pol** : L'identifiant de la police d'assurance, permettant de lier ces données avec d'autres bases de données comme contract_data et telematics_data.
- **NB_Claim** : Le nombre de réclamations déposées par le détenteur de la police pendant la période observée. Cette information est cruciale pour analyser la fréquence des sinistres.
- **AMT_Claim**: Le montant total des réclamations déposées par le détenteur de la police. Ce montant est essentiel pour évaluer la sévérité des sinistres et leur impact financier.

2. contract_data (DB_CNT.xlsx)

Cette base de données est susceptible de contenir des informations relatives aux contrats d'assurance souscrits par les clients. Voici les types de variables typiquement incluses :

- **Id_pol** : L'identifiant unique de la police d'assurance, servant de clé principale.
- **Duration** : La durée de couverture de l'assurance, souvent exprimée en jours.
- **Insured.age** : L'âge de l'assuré, un facteur important dans le calcul du risque et de la prime.
-**Insured.sex** : Le sexe de l'assuré, qui peut influencer les statistiques de risque.
- **Car.age** : L'âge du véhicule assuré, plus le véhicule est vieux, potentiellement plus le risque de panne est élevé.
- **Marital** : Le statut marital de l'assuré, pouvant affecter le profil de risque.
- **Car.use** : L'utilisation du véhicule, comme privée ou commerciale, affectant également le risque.
- **Credit.score** : Le score de crédit de l'assuré, indicatif de sa fiabilité financière.
- **Region** : La région de résidence de l'assuré, car les risques varient géographiquement.
- **Annual.miles.drive** : Les miles annuels prévus, affectant directement l'exposition au risque.
- **Years.noclaims** : Le nombre d'années sans réclamation, un indicateur de la prudence du conducteur.
- **Territory** : Le code territorial où le véhicule est principalement utilisé.

3. telematics_data (DB_TELEMATICS.csv)

Cette base de données regroupe des informations issues de la télémétrie, reflétant le comportement de conduite des assurés. Les variables couramment trouvées incluent :

- **Id_pol** : Identifiant unique pour chaque police d'assurance.
- **Annual.pct.driven** : Pourcentage annuel du temps passé sur la route.
- **Total.miles.driven** : Total des miles parcourus, indicatif de l'utilisation globale du véhicule.

- **Pct.drive.xhrs, Pct.drive.xxx** : Pourcentage de conduite pendant certaines heures ou certains jours, utile pour évaluer les habitudes de conduite.
- **Pct.drive.rushxx** : Pourcentage de conduite pendant les heures de pointe, important pour évaluer le risque d'accidents.
- **Avgdays.week** : Nombre moyen de jours de conduite par semaine.
- **Accel.xxmiles, Brake.xxmiles** : Incidences des accélérations et freinages brusques, indicateurs de comportements de conduite agressifs ou dangereux.
- **Left.turn.intensityxx, Right.turn.intensityxx** : Intensité des virages à gauche et à droite, qui peut être reliée aux techniques de conduite et aux risques d'accidents.

Ces données sont utilisées pour évaluer le risque basé sur les habitudes de conduite, crucial pour la tarification et la gestion des risques dans les assurances automobiles basées sur l'usage (UBI - Usage-Based Insurance). Chaque ensemble de données joue un rôle clé dans la compréhension complète du profil de risque des assurés, et leur combinaison permet une analyse exhaustive nécessaire pour la modélisation prédictive en assurance.

B- PRETRAITEMENT DES DONNEES ET EDA

B.1 - Claims Data

La base de données claims_data, issue de DB_SIN.txt, contient des informations détaillées sur les réclamations déposées par les assurés entre 2013 et 2016. Les variables principales présentes sont :

- Id_pol : Identifiant unique de la police d'assurance.
- NB_Claim : Nombre de réclamations déposées par l'assuré.
- AMT_Claim : Montant total des réclamations déposées.

Une première exploration statistique via describe() a fourni un résumé des principales statistiques des variables :

- **NB_Claim** :
  - Nombre d'Observations : 4309.
  - Moyenne : 1.05.
  - Écart-Type : 0.22.
  - Minimale : 1.
  - 25ème Percentile : 1.
  - Médiane (50ème Percentile) : 1.
  - 75ème Percentile : 1.
  - Maximale : 3.

La majorité des polices d'assurance ont été associées à une seule réclamation (75 % des données). Le maximum de trois réclamations indique une faible répétition des sinistres pour une même police, suggérant que la plupart des sinistres sont uniques.

- **AMT_Claim** :

- Nombre d'Observations : 4272.
- Moyenne : 3183.90.
- Écart-Type : 5160.61.
- Minimale : 0.
- 25ème Percentile : 522.87.
- Médiane (50ème Percentile) : 1713.14.
- 75ème Percentile : 3742.02.
- Maximale : 104074.89.

Le montant moyen des réclamations s'élève à 3183.90, avec un écart-type significatif reflétant une grande variabilité des montants. La médiane (1713.14) étant nettement inférieure à la moyenne, cela indique une distribution asymétrique avec une majorité de valeurs plus basses. Le montant maximal de 104074.89 suggère que quelques réclamations sont très élevées, tirant la moyenne vers le haut.

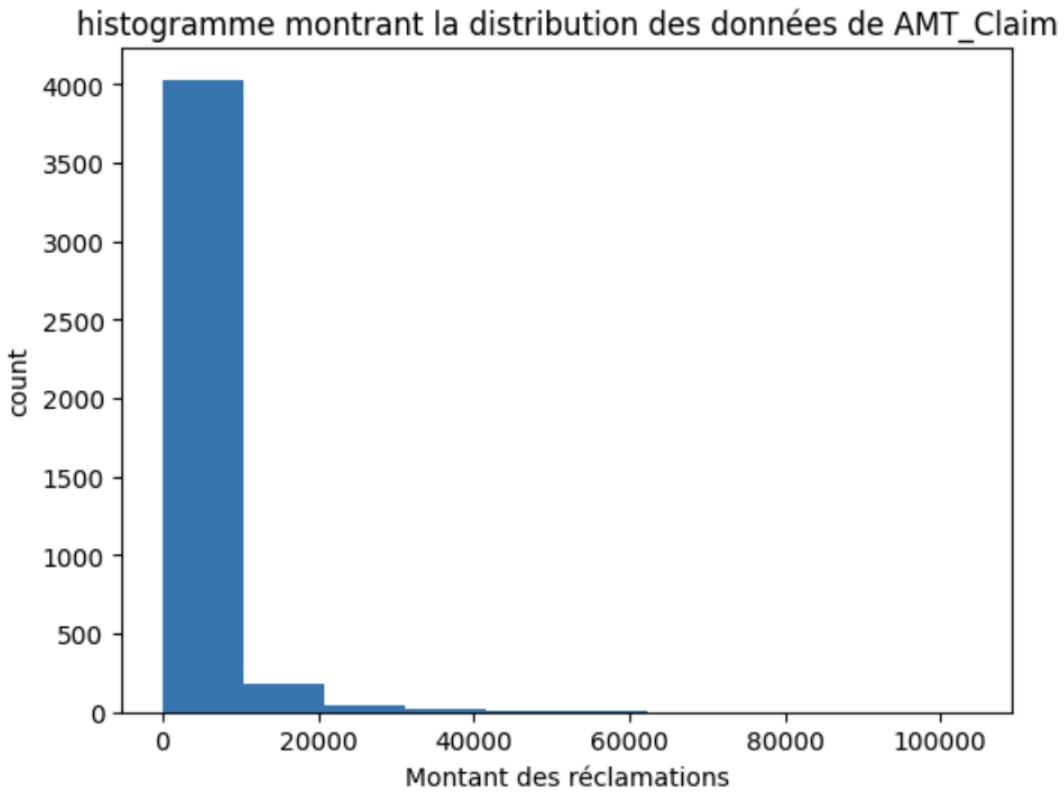
Nettoyage et Uniformisation des Données

Les valeurs de NB_Claim ont été converties en entier après le nettoyage des annotations textuelles.
Les virgules ont été remplacées par des points dans AMT_Claim, permettant la conversion en flottant.
Les valeurs 'ANN' dans AMT_Claim ont été remplacées par NaN pour une gestion appropriée des valeurs manquantes.
Les doublons dans claims_data ont été supprimés.
Les valeurs aberrantes dans AMT_Claim au-delà du troisième quartile + 1,5 IQR ont été retirées, représentant environ 8 % des observations totales.

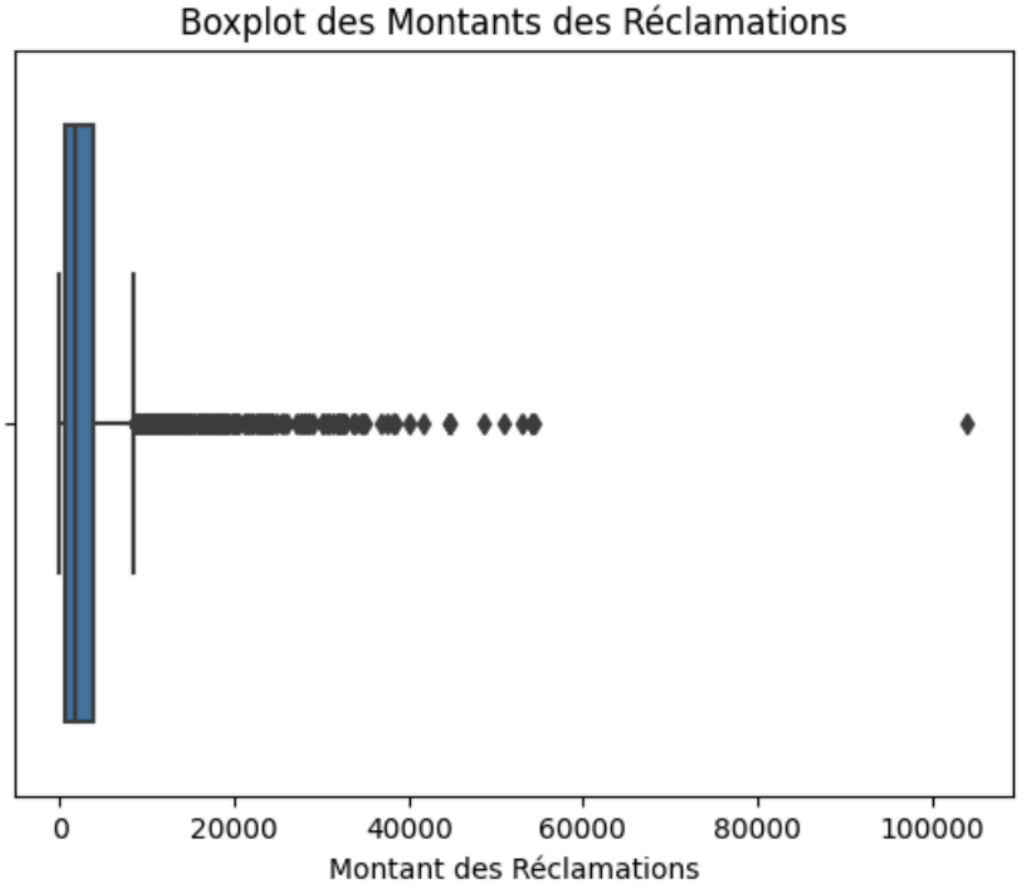
Analyse Graphique

- **Distribution du Montant des Réclamations :**

Nous constatons grâce à l'histogramme que la grande majorité des réclamations dans 'claims_data' a un montant compris entre 0 et 20000. L'observation d'une majorité de montants de réclamations se regroupant entre 0 et 20,000 pourrait suggérer une distribution gamma, étant donné que cette distribution est souvent utilisée pour modéliser des données qui sont fortement asymétriques et limitées à des valeurs positives. La forme de la distribution gamma correspond bien à cette concentration de valeurs basses avec une longue queue vers des valeurs plus élevées.



- **Boîte à Moustache : Montre la présence de valeurs aberrantes, justifiant leur retrait.**

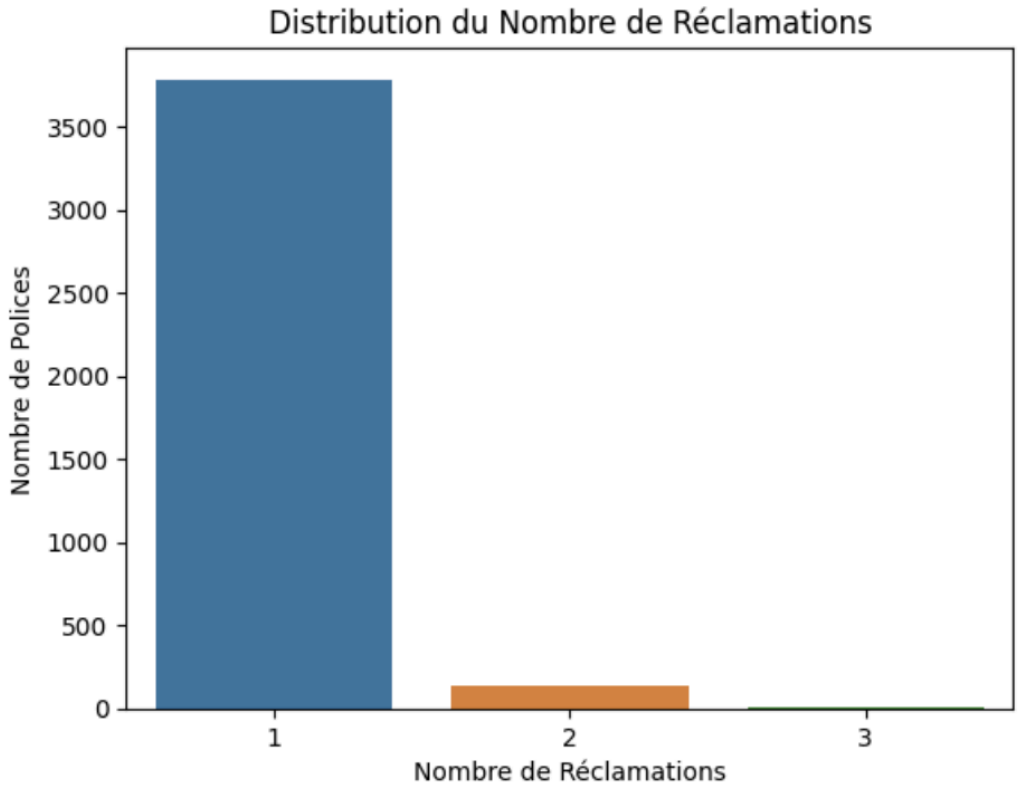


Le Boxplot suggère ici la présence d'un nombre assez élevé de valeurs aberrantes dans notre df. Nous allons calculer les quantiles Q1 et Q3 afin de déterminer l'écart interquartile(IQR). Vu que le montant des sinistres est supérieur ou égal à 0,sont considérées comme valeurs aberrantes, les valeurs supérieures à Q3+1,5*IQR.

Dans notre contexte, une valeur aberrante (ou outlier) désigne un montant de réclamation qui s'écarte significativement de la majorité des autres montants dans les données, en étant beaucoup plus élevé que ce que l'on observe typiquement. Ces valeurs peuvent indiquer des sinistres exceptionnellement coûteux ou des erreurs de saisie.

Retirer les valeurs aberrantes peut être nécessaire pour éviter que ces valeurs extrêmes ne biaisent les analyses statistiques et les modèles prédictifs, améliorant ainsi l'exactitude et la robustesse des résultats. De plus, l'élimination des outliers permet d'obtenir une meilleure représentation des tendances générales et des comportements typiques dans les données.

- **Distribution du Nombre de Réclamations** :



Sur la période allant de 2013 à 2016, plus de 75 % des polices d'assurance n'ont eu qu'une seule réclamation, confirmant une tendance claire dans la fréquence des sinistres.

Les analyses descriptives et graphiques de claims_data révèlent des tendances clés dans les comportements de réclamation. La grande majorité des polices d'assurance sont associées à une seule réclamation, principalement concentrée dans des montants faibles. Les valeurs aberrantes ont été gérées de manière appropriée, laissant une base de données propre et exploitable pour les analyses ultérieures.

B.2 - Prétraitement des Données et Analyse Exploratoire (EDA) de Contract Data

La base de données contract_data, issue de DB_CNT.xlsx, contient des informations relatives aux contrats d'assurance souscrits par les clients. Voici un aperçu des variables principales présentes :

- Id_pol : Identifiant unique de la police d'assurance, clé principale.
- Duration : Durée de la couverture en jours.
- Insured.age : Âge de l'assuré.
- Insured.sex : Sexe de l'assuré.
- Car.age : Âge du véhicule.
- Marital : Statut marital de l'assuré.
- Car.use : Utilisation du véhicule (commercial, privé, etc.).
- Credit.score : Score de crédit de l'assuré.
- Region : Région de résidence de l'assuré.
- Annual.miles.drive : Miles annuels prévus.
- Years.noclaims : Nombre d'années sans réclamation.
- Territory : Code territorial où le véhicule est principalement utilisé.

1. Statistiques Descriptives

L'exploration statistique initiale de contract_data a révélé :

- Insured.sex :
  - Valeurs Uniques : 2.
  - Valeur la plus Fréquente : Male.
  - Fréquence : 53883.

- Marital :
  - Valeurs Uniques : 2.
  - Valeur la plus Fréquente : Married.
  - Fréquence : 69763.

- Car.use :
  - Valeurs Uniques : 4.
  - Valeur la plus Fréquente : Commute.
  - Fréquence : 49771.

- Region :
  - Valeurs Uniques : 2.
  - Valeur la plus Fréquente : Urban.
  - Fréquence : 77958.

2. Nettoyage et Prétraitement

1. Filtrage des Catégories Pertinentes :
   - La variable Car.use comportait des catégories peu fréquentes et erronées. Nous avons retenu uniquement les catégories principales (Private, Commute, Farmer, Commercial).

2. Standardisation des Valeurs :
  - Les valeurs de Insured.sex et Marital ont été uniformisées pour réduire les variations linguistiques ou textuelles :
    - H a été remplacé par Male et F par Female.
    - Marié a été remplacé par Married et Celib par Single.

3. Gestion des Valeurs Manquantes :
  - Certaines lignes de Marital et Region comportaient des valeurs manquantes, remplacées par les valeurs les plus fréquentes.

4. Gestion des Doublons :
  - Les doublons purs ont été supprimés, réduisant la taille initiale de contract_data.
  - Les doublons basés sur Id_pol ont été conservés si des changements de statut marital étaient détectés.

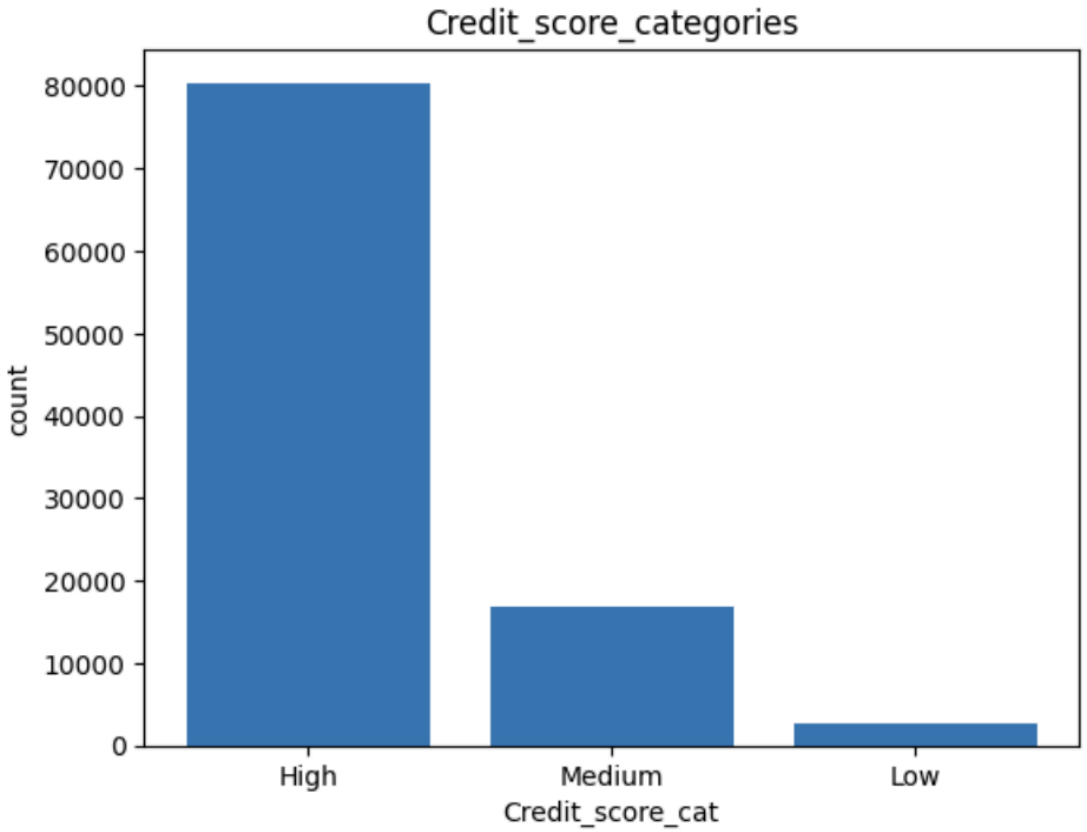
La majorité des assurés sont des hommes mariés vivant en milieu urbain (31,44 %), suivis par les femmes mariées en milieu urbain (22,22 %).

Les analyses de contract_data montrent que certaines variables présentent une grande stabilité, notamment l'âge et le sexe des assurés. D'autres variables, telles que Duration, changent plus fréquemment, indiquant des variations dans la durée de couverture. Ces observations fournissent un aperçu utile pour comprendre les comportements des assurés et anticiper leurs besoins lors de la modélisation.

Exploration Analytique et Graphique

Binning et Catégorisation du Score de Crédit

Nous avons regroupé les scores de crédit en trois catégories : Low, Medium, et High.
Les scores ont été divisés en utilisant des intervalles uniformes entre la valeur minimale (422) et maximale (900).
  - Résultat de la distribution :
    - High : 80 422 assurés.
    - Medium : 16 700 assurés.
    - Low : 2696 assurés.

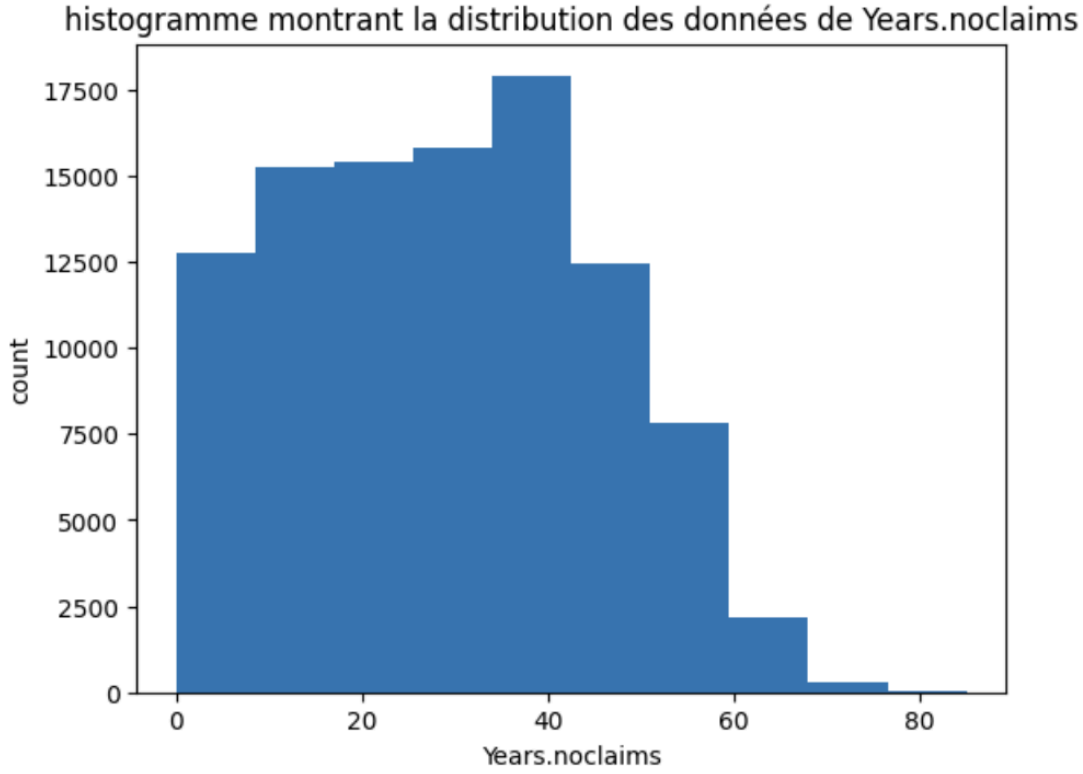


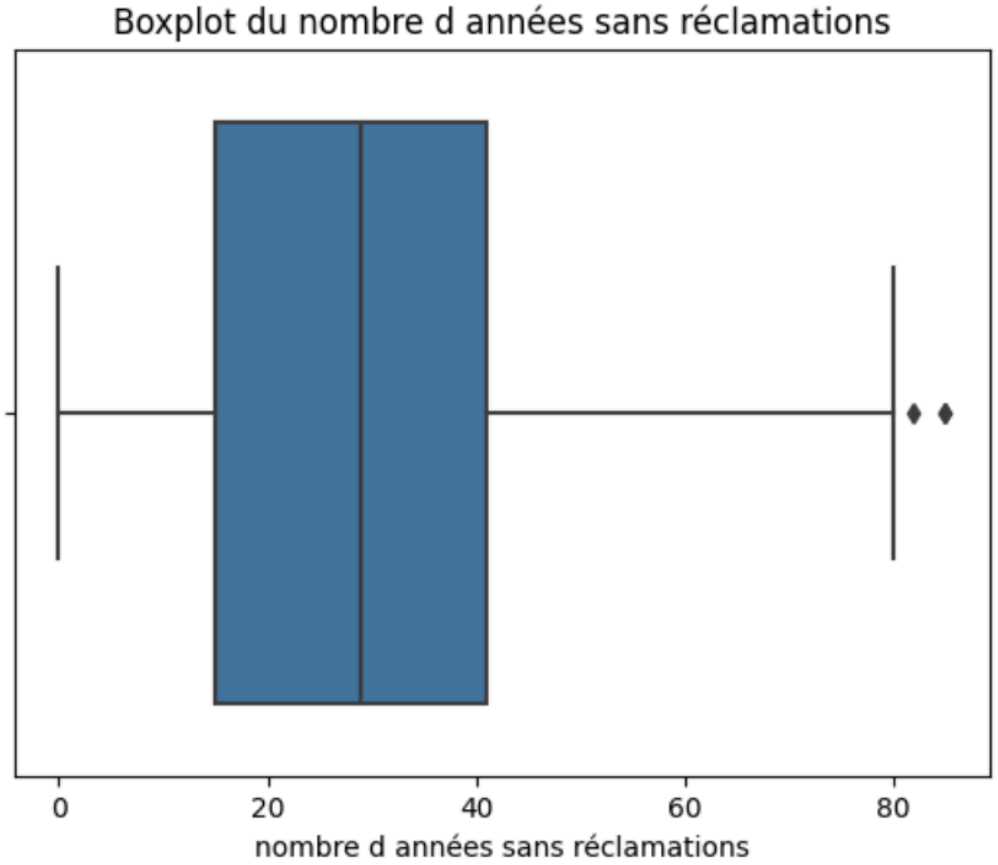
Un graphique à barres illustre la distribution des scores de crédit par catégorie. La majorité des assurés se situent dans la catégorie High, avec un écart marqué par rapport aux catégories inférieures.

Distribution des Years.noclaims

L'histogramme montre que la plupart des assurés ont entre 0 et 40 années sans réclamation, avec un pic à 40. Cette distribution suggère une majorité de conducteurs prudents et expérimentés.

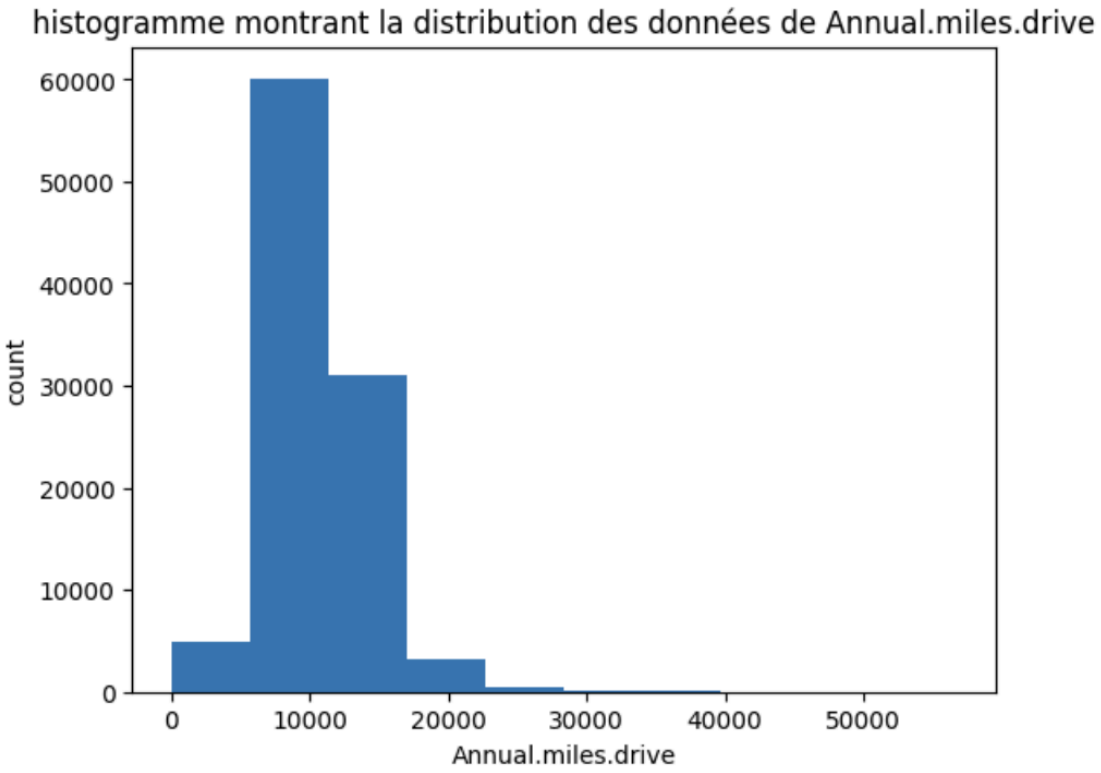
La Boîte à Moustache (Boxplot) indique très peu de valeurs aberrantes dans la distribution des années sans réclamation.


histogramme montrant la distribution des données de Years.noclaims
count
17500
15000
12500
10000
7500
5000
2500
0
0
20
40
60
80
Years.noclaims

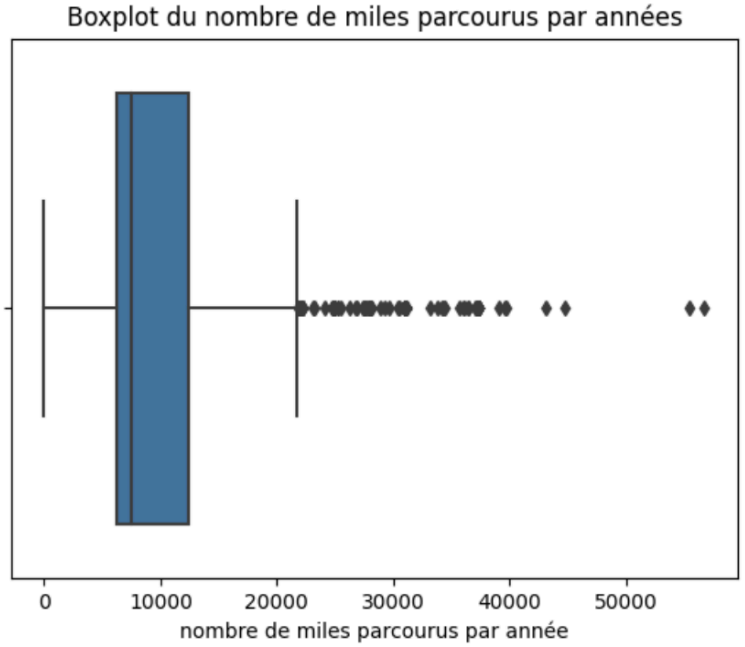

Boxplot du nombre d années sans réclamations
0
20
40
60
80
nombre d années sans réclamations

Distribution de Annual.miles.drive

La majorité des assurés conduisent entre 0 et 20 000 miles par an.
La distribution montre un pic autour de 10 000 miles.



La boîte à moustache révèle certaines valeurs aberrantes au-delà de 20 000 miles, représentant 0,64 % des données totales.
Les outliers ont été retirés pour ne conserver que les valeurs pertinentes.



Gestion des Valeurs Négatives dans Car.age

1919 valeurs négatives ont été trouvées, représentant 1,93 % du total. Ces valeurs sont probablement dues à des erreurs de saisie.
Les lignes avec des valeurs négatives ont été supprimées, ne laissant plus de valeurs aberrantes dans Car.age.

Statistiques Descriptives Finales

Après ces étapes de nettoyage et de catégorisation :

- Duration : Moyenne de 314 jours.
- Insured.age : Moyenne de 51 ans.
- Car.age : Moyenne de 5,6 ans.
- Credit.score : Moyenne de 801.
- Annual.miles.drive : Moyenne de 9007 miles.
- Years.noclaims : Moyenne de 28 ans.

La base contract_data peut maintenant servir pour des analyses plus approfondies et la modélisation prédictive.

B.3 - Prétraitement des Données et Analyse Exploratoire de Telematics Data

La base de données télématique est une source riche en informations sur le comportement de conduite des assurés. Elle comprend des variables détaillées, allant du pourcentage de conduite sur une base annuelle aux habitudes spécifiques d'accélération, de freinage et de conduite aux heures de pointe. L'analyse approfondie de ces données offre des insights précieux sur les profils de risque des conducteurs, aidant les assureurs à affiner leurs modèles prédictifs et leurs stratégies de tarification.

Avant de procéder à l'analyse, une série d'étapes de nettoyage et de transformation des données a été effectuée pour garantir l'uniformité, la fiabilité et la pertinence des variables.

Uniformisation et Conversion des Types de Données

La colonne Id_pol contenait un préfixe cnt_ qui a été supprimé afin de rendre l'identifiant compatible avec d'autres jeux de données. Toutes les colonnes ont été initialement converties en chaînes de caractères, pour permettre la manipulation des formats textuels. Les virgules ont été remplacées par des points pour standardiser les décimales. Les colonnes numériques ont ensuite été converties en float à l'exception de Id_pol, qui reste un identifiant unique.

Gestion des Valeurs Nulles et Doublons

Nous avons supprimé les lignes où toutes les colonnes numériques étaient nulles ou égales à zéro, car elles ne contenaient aucune donnée utile. En termes de doublons, la base ne comportait aucune répétition exacte d'une ligne.

Analyse Exploratoire et Réduction de la Dimensionnalité

Une analyse exploratoire a ensuite été menée pour identifier les relations entre les différentes variables. La matrice de corrélation, visualisée sous forme de heatmap, a révélé plusieurs relations positives fortes entre les variables reflétant des comportements similaires. Cela a mis en évidence l'utilité d'une analyse en composantes principales (PCA) pour réduire la dimensionnalité du jeu de données tout en conservant l'information essentielle.

PCA pour la Réduction de la Dimensionnalité

Nous avons effectué une PCA en limitant le nombre de composantes à cinq, permettant de conserver une grande partie de la variance tout en réduisant la complexité. Le DataFrame résultant comprend ces cinq composantes principales et l'identifiant Id_pol. Chaque composante représente un groupe de comportements liés, comme le montre la variance expliquée par chaque composante.

- PC1 : L'intensité annuelle de conduite est capturée par Annual.pct.driven, représentant la proportion annuelle de temps passé à conduire.
- PC2 : Les comportements de virage sont saisis par Left.turn.intensity et Right.turn.intensity, qui indiquent la fréquence des virages à forte intensité.
- PC3 : Les actions d'accélération et de freinage sont regroupées pour capturer les comportements agressifs ou prudents.
- PC4 : Les manœuvres de freinage et de virages à droite fournissent des informations sur les actions plus subtiles.
- PC5 : Cette composante reflète des variations plus subtiles dans les habitudes de conduite.

Chaque composante principale représente un aspect distinct du comportement de conduite. Leur utilisation combinée permet aux assureurs d'obtenir une vue plus claire des habitudes des conducteurs.

Les habitudes de conduite capturées dans ces composantes permettent de distinguer plus facilement les conducteurs prudents de ceux plus enclins à des comportements risqués. Les assureurs peuvent ainsi identifier des profils spécifiques, permettant une meilleure segmentation des assurés.

En fournissant des indicateurs synthétiques et moins corrélés, les composantes principales simplifient les modèles prédictifs, les rendant plus robustes. Elles peuvent également être intégrées aux algorithmes de tarification pour affiner les prévisions.

L'analyse exploratoire des données télématiques et l'application de la PCA offrent une perspective unique sur les comportements de conduite. Ces méthodes de réduction de la dimensionnalité permettent non seulement de simplifier les relations complexes entre les variables, mais aussi de mettre en lumière les facteurs déterminants des habitudes de conduite. Ces insights sont essentiels pour améliorer les modèles prédictifs de sinistres et ajuster les stratégies tarifaires.

B.4- Synthèse des Analyses Précédentes

1. Statistiques Descriptives:

Les analyses descriptives des bases de données individuelles ont révélé des tendances importantes :

contract_data :
- Les conducteurs mariés représentent la majorité des assurés, principalement en milieu urbain.
- Les scores de crédit élevés prédominent, indiquant des profils financiers globalement stables.
- La distribution du kilométrage annuel montre une majorité de conducteurs parcourant moins de 20 000 miles.

df_telematics :
- Les habitudes de conduite révèlent une majorité de conducteurs prudents.
- Les comportements d'accélération et de freinage brusques sont rares, reflétant une conduite généralement sûre.
- Les habitudes de virage diffèrent en intensité, aidant à segmenter les conducteurs en profils distincts.

claims_data :
- La plupart des contrats d'assurance ont enregistré une seule réclamation, ce qui montre que les réclamations multiples sont rares.
- Les montants de réclamation varient considérablement, la majorité étant inférieure à 20 000 unités monétaires.
- Les montants de réclamations extrêmes ont été identifiés et filtrés pour éviter de fausser les analyses.

2. Prétraitement et Gestion des Anomalies:

- Valeurs Manquantes : Les valeurs manquantes dans les différentes bases ont été gérées par remplacement par la modalité la plus fréquente ou par suppression des enregistrements non pertinents.
- Doublons : Les doublons ont été identifiés et supprimés pour garantir la précision des analyses.

3. Analyse en Composantes Principales (PCA):
- La PCA a permis de réduire la dimensionnalité des données télématiques tout en conservant les relations complexes entre les variables.
- Les cinq composantes principales capturent les comportements de conduite et facilitent la modélisation prédictive.

4. Fusion des Bases de Données:
- Les bases contract_data et df_telematics ont été fusionnées pour créer une base enrichie, df_insurance.
- La fusion de df_insurance avec claims_data a permis de constituer une base de données finale, claims_insurance, regroupant les informations clés sur les contrats, les comportements de conduite et les réclamations.

La consolidation de ces ensembles de données nous offre une vue complète du profil des assurés. Les informations combinées sur les comportements de conduite, les caractéristiques des contrats et les réclamations permettent de créer des modèles prédictifs robustes, d'améliorer la segmentation des clients et de perfectionner les stratégies tarifaires.

C - Économétrie

Dans cette section, nous appliquons différentes méthodes économétriques pour comprendre les facteurs qui influencent les sinistres dans le secteur de l'assurance. La variable cible principale de cette étude est AMT_Claim, le montant total des sinistres, car c'est une mesure essentielle du risque financier que l'assureur doit couvrir. Nous cherchons à identifier les caractéristiques significatives qui expliquent la variabilité de cette variable. Voici une description détaillée du processus d'analyse.

C.1 - Modèle de Régression Linéaire Multiple

Objectif: Identifier les facteurs qui influencent le montant total des sinistres (AMT_Claim).

Justification: La régression linéaire multiple permet d'évaluer les effets de plusieurs variables explicatives sur une variable cible continue. C'est une méthode classique pour estimer les coefficients de chaque prédicteur tout en contrôlant les autres variables.

Approche:
1. Préparation des Données:
- Filtrage et Encodage:
  - Nous avons supprimé les colonnes corrélées et peu significatives, comme Insured.age.
  - Les variables catégorielles ont été encodées via OneHotEncoder pour créer des variables binaires (dummy variables).
- Sélection des Variables:
  - Nous avons conservé des prédicteurs pertinents, tels que Duration, Years.noclaims, Region, Credit_score_cat.

2. Construction du Modèle:

- La régression linéaire multiple a été réalisée en utilisant sm.OLS de la bibliothèque statsmodels.
- AMT_Claim est la variable dépendante, et toutes les autres variables sélectionnées ont été ajoutées comme prédicteurs, avec une constante (intercept).

3. Évaluation du Modèle:

- Coefficients Significatifs:
  - Duration: Une augmentation d'une unité dans la durée d'assurance correspond à une augmentation moyenne de 3,08 dans le montant des sinistres, ce qui est statistiquement significatif ($p < 0,001$).
  - Years.noclaims: Chaque année sans sinistre réduit en moyenne le montant des réclamations de 14,24, ce qui est également significatif ($p < 0,001$).
  - Marital_Single: Être célibataire est associé à une augmentation de 172,87 dans le montant des sinistres.
  - Car.use_Commute: Utiliser la voiture pour le trajet domicile-travail augmente en moyenne le montant des sinistres de 479,71.
  - Credit_score_cat_Low: Les assurés ayant un faible score de crédit présentent des sinistres plus élevés en moyenne (+992,22) par rapport à ceux ayant un score élevé.

- Coefficients Non Significatifs:
  - Plusieurs variables (par exemple, Principal Component (1-5), Car.age, Territory) n'ont pas d'effet significatif sur AMT_Claim, suggérant qu'elles n'apportent pas d'informations utiles à la prédiction.

- Multicolinéarité:
  - Les VIF (Variance Inflation Factors) ont été calculés pour chaque variable explicative.
  - Les variables Car.use_Commute et Car.use_Private présentent des VIF élevés, indiquant une possible multicolinéarité.

- Performances du Modèle:
  - Le R-carré du modèle est de 0,059, ce qui montre que le modèle n'explique que 5,9 % de la variance totale du montant des sinistres.
  - Le RMSE (Root Mean Squared Error) de 1925,69 indique un écart moyen élevé entre les valeurs prédites et réelles.

C.2 - Modèle Polynomial

Objectif: Examiner les interactions complexes entre les prédicteurs en utilisant un modèle polynomial.

Justification: Un modèle polynomial permet de capturer les interactions non linéaires entre les prédicteurs, offrant une meilleure compréhension des relations complexes.

Approche:
1. Transformation Polynomiale:
- Les données ont été transformées en utilisant PolynomialFeatures de degré 2.
- Cela crée des termes quadratiques et d'interaction entre les variables.

2. Construction du Modèle:
   - Un modèle de régression linéaire a été construit en utilisant ces nouvelles variables transformées.

3. Évaluation du Modèle:
   - RMSE: Le RMSE du modèle polynomial est de 19 459, suggérant que ce modèle n'explique pas suffisamment la variance.
   - Coefficients: Les coefficients indiquent des valeurs élevées, parfois négatives, suggérant un surajustement.

C.3 - Conclusion des Analyses Économétriques

Les analyses montrent que certaines variables ont un impact significatif sur AMT_Claim, mais le faible R-carré du modèle linéaire multiple indique que d'autres facteurs restent à découvrir. Les modèles polynomiaux souffrent d'un surajustement, suggérant que ces interactions ne sont peut-être pas pertinentes pour ce cas.

On pourrait envisager de:
1. Réduire la Multicolinéarité:
   - Grouper certaines catégories ou supprimer les variables fortement corrélées.

2. Explorer d'Autres Modèles:
   - Régression Ridge ou Lasso pour la régularisation.
   - Random Forests ou XGBoost pour capturer les interactions complexes.

Les analyses futures peuvent améliorer les modèles existants et trouver de nouvelles relations pour mieux comprendre et prédire les sinistres.

CONCLUSION

L'analyse des données d'assurance, allant du nettoyage des bases de données à l'application des techniques économétriques, révèle les nombreuses complexités et subtilités de ce domaine. Nous avons soigneusement parcouru les étapes du traitement des données, de l'exploration graphique et descriptive à la modélisation statistique, afin d'identifier les facteurs qui influencent les montants des sinistres.

1. Data Cleaning:
   - Prétraitement: Les différentes bases de données ont été nettoyées pour éliminer les doublons, corriger les anomalies et combler les valeurs manquantes. Cela a permis de garantir une cohérence des variables clés pour la modélisation ultérieure.
   - Fusion des Bases de Données: La fusion des ensembles de données a été réalisée en tenant compte des liens entre les polices d'assurance, les comportements de conduite et les sinistres. Cette intégration offre une vue globale des caractéristiques du contrat et du profil de risque.

2. Exploration Graphique et Descriptive:
   - contract_data: La majorité des assurés sont mariés, vivent en milieu urbain, et ont un score de crédit élevé, indiquant un profil financier relativement stable.
   - df_telematics: Les habitudes de conduite montrent une majorité de conducteurs prudents, mais certaines variables comme les virages et les freinages brusques permettent de différencier les comportements.
   - claims_data: La plupart des contrats sont associés à un seul sinistre, généralement de faible montant, mais quelques cas extrêmes faussent la distribution.

3. Analyse Économétrique:
   - Régression Linéaire Multiple: Cette méthode a révélé que la durée d'assurance, les années sans sinistre, et le score de crédit faible sont les principaux facteurs influençant le montant des sinistres. Cependant, le faible R-carré suggère que d'autres variables restent à découvrir.
   - Modèle Polynomial: Les interactions non linéaires entre les variables ont été explorées via un modèle polynomial, mais celui-ci a souffert d'un surajustement.

Le travail combiné de nettoyage, d'exploration et de modélisation a permis de comprendre les principales tendances et les interactions clés dans les données d'assurance. Bien que certaines variables explicatives montrent des associations significatives, une meilleure compréhension du phénomène des sinistres nécessitera des analyses plus approfondies. Ces résultats ouvrent la voie à des modèles prédictifs plus robustes et à des stratégies tarifaires optimisées pour les assureurs.