EXPLORATION DES DONNÉES

1 Description du jeu de données

La base de données à notre disposition contient 90 000 déclarations de sinistres `Arrêt Travail` en Australie et générés de manière synthétique. Pour chacune des observations contenu dans la base, nous avons la synthèse de 3 catégories d'informations :

- les informations contractuelles relatives au travailleur et à la nature de son contrat de travail ;
- les informations concernant la personnes intervenant dans le contrat tel que le statut matrimonial, le sexe, le nombre de personne à charge ;
- les informations textuelles décrivant la survenance du sinistre.

skimpy summary

Data Summary

dataframe	Values
Number of rows	54000
Number of columns	16

Data Types

Column Type	Count
int32	6
string	5
float64	3
datetime64	2

number

column_name	NA	NA %	mean	sd	p0	p25	p50	p7
Age	0	0	33.84	12.12	13	23	32	
DependentChildren	0	0	0.1192	0.5178	0	0	0	
DependentsOther	0	0	0.009944	0.1093	0	0	0	
WeeklyWages	0	0	416.4	248.6	1	200	392.2	5
HoursWorkedPerWeek	0	0	37.74	12.57	0	38	38	
DaysWorkedPerWeek	0	0	4.906	0.5521	1	5	5	
InitialIncurredCalims Cost	0	0	7841	20580	1	700	2000	95
UltimateIncurredClaim Cost	0	0	11000	33390	121.9	926.3	3371	81
AnneAccident	0	0	1997	5.188	1988	1992	1997	20

datetime

column_name	NA	NA %	first	last
DateTimeOfAccident	0	0	1988-01-01 09:00:00	2005-12-31 10:0
DateReported	0	0	1988-01-08	2006-09-23

string

column_name	NA	NA %	words per row	
ClaimNumber	0	0	1	
Gender	0	0	1	
MaritalStatus	29	0.05	1	
PartTimeFullTime	0	0	1	
ClaimDescription	0	0	7	

End

2 Qualité des données

Hormis de la variable `MaritalStatus` renseignant sur le statut matrimonial avec 5% de données manquantes, les variables sont bien remplies.

- Nous avons dans la base des travailleurs de plus de 75 ans, ce qui est pour le moins atypique.
- Les sinistres de notre base concernent les survenances de 1988 à 2005.
- Le portefeuille est majoritairement jeune, avec un âge médian de 32 ans.
- Les sinistrés n'ont globalement pas de personne à charge, nous pouvons "binariser" cette information.
- En moyenne, l'assureur estime le coût initial du sinistre à 7776, avec un coût médian de 2000.
- Le salaire hebdomadaire médian de chaque sinistré est de 393.3.

2.1 Analyse en cascade des données

Overview

Brought to you by YData (https://ydata.ai/?utm_source=opensource&utm_medium=ydataprofiling&utm_campaign=report)

Dataset statistics

Number of variables	16
Number of observations	54000
Missing cells	29
Missing cells (%)	< 0.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	6.4 MiB
Average record size in memory	124.0 B

Variable types

Text	2
DateTime	2
Numeric	8
Categorical	4

Alerts

2.2 Correction des données

La variable `Gender` renseignant sur le sexe a des observations où le sexe est inconnu : nous les remplaçons par le mode.

La variable `MaritalStatus` renseignant sur le statut matrimonial a une proportion non-négligeable de valeurs inconnues : nous la laissons en l'état, car elle peut représenter une catégorie non prise en compte comme les concubins ou les pacsés.

Nous avons "binarisé" la présence de personne dépendante à la charge du sinistré.

Nous avons également borné le nombre d'enfants à 4 au vu du faible nombre de valeur au-delà.

La variable `HoursWorkedPerWeek` renseignant sur le nombre d'heures de travail par semaine a des valeurs aberrantes. Selon le site https://www.studyrama.com/, légalement, le temps de travail hebdomadaire en Australie est fixé à 40 heures par semaine. En pratique, il varie entre 35 et 40 heures (le plus souvent 38 heures) selon les contrats. Au-delà des heures légales, les heures supplémentaires sont comptées 50 % plus chères. Par contre, il n'y a pas en Australie de limite maximale spécifique pour les heures supplémentaires. Nous fixons arbitrairement le nombre d'heures supplémentaires à 20 heures par semaine et corrigeons les données.

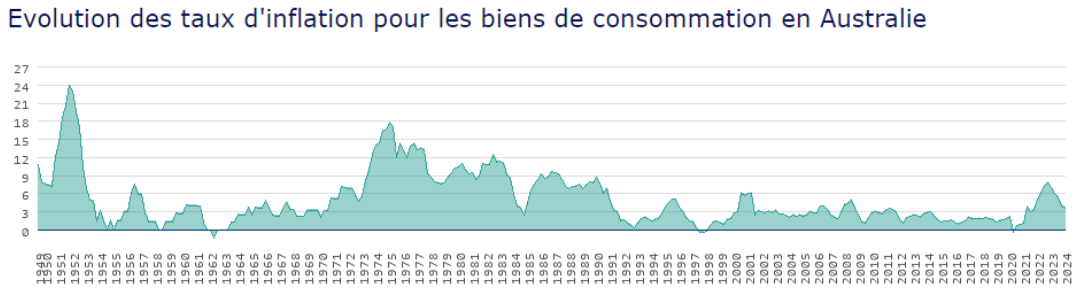
Nous remplaçons les salaires hebdomadaires de moins de 5 par la médian et l'estimation initiale de moins de 50 par la médiane.

3 Prise en compte de l'inflation et enrichissement de la base

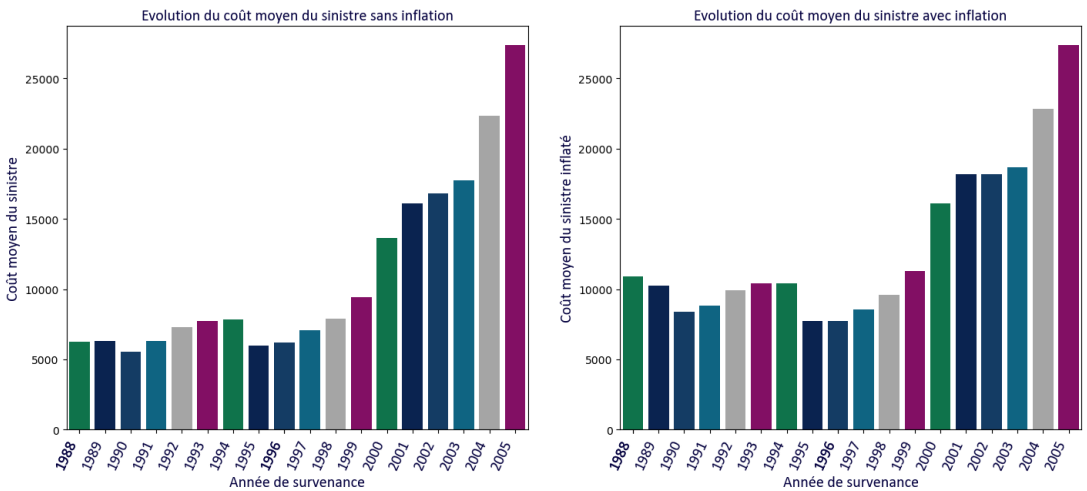
Pour prendre en compte l'inflation dans le coût des sinistres d'une année à l'autre, il est important d'ajuster les coûts des sinistres à une valeur constante pour que les comparaisons soient pertinentes. Ainsi, pour une année de référence t_0, le coût ajusté du sinistre est défini par :

$Coût\_ajt_t = Coût_t * \frac{IPC_{t_0}}{IPC_t}$, où $IPC_t = IPC_{t-1} * (1 + taux\_inflation)$.

Le graphique ci-dessous présente l'évolution de l'inflation en Australie pour les 75 dernières années.

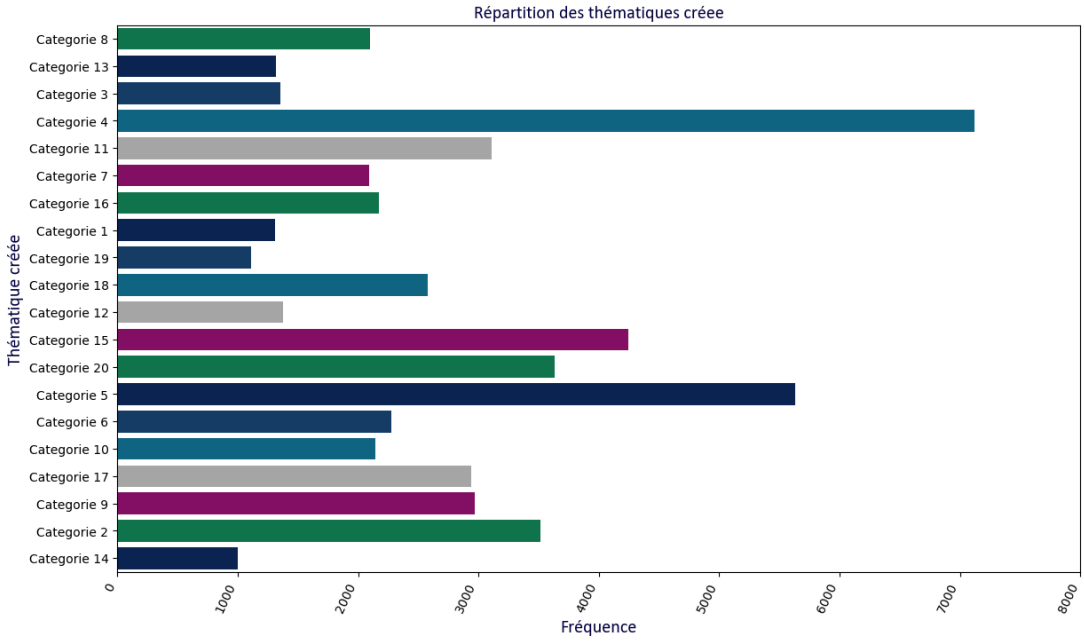


Nous représentons ensuite l'évolution du coût moyen de sinistre avant la prise en compte de l'inflation et après la prise en compte de l'inflation.



4 Analyse de la description des sinistres.

A partir du texte non structuré issue de la description du sinistre, nous allons utiliser la modélisation automatique de sujet permettant de détecter les sujets latents abordé dans un corpus de documents pour ensuite assigner les sujets détecté à ces corpus de documents : c'est donc de l'apprentissage non-supervisé permettant de créer des catégories de sujets. La méthode LDA (Latent Dirichlet Allocation) est la plus ancienne et la plus populaire.



Après correction des données et prise en compte de l'inflation, nous obtenons la nouvelle analyse en cascade ci-dessous.

Overview

Brought to you by YData (https://ydata.ai/?utm_source=opensource&utm_medium=ydataprofiling&utm_campaign=report)

Dataset statistics

Number of variables	18
Number of observations	54000
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	6.8 MiB
Average record size in memory	132.0 B

Variable types

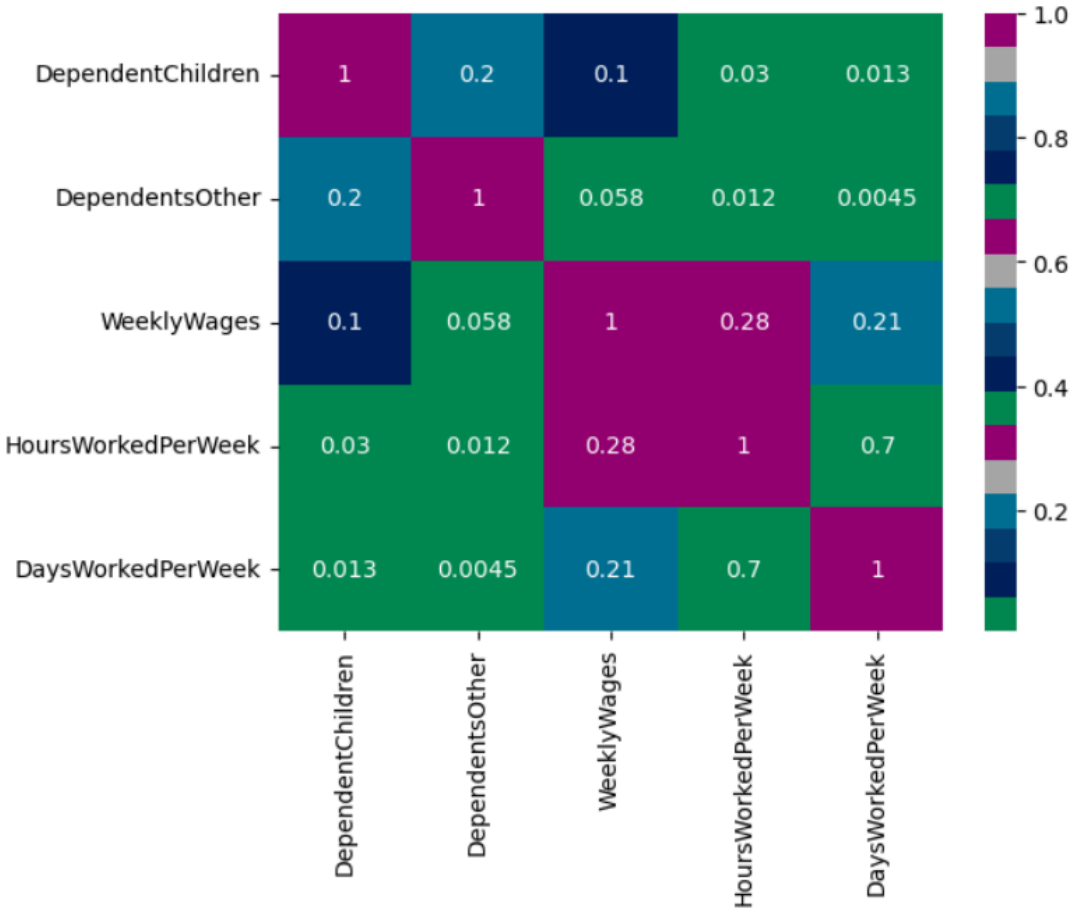
Numeric	11
Categorical	7

Alerts

`ClaimInflate` is highly overall correlated with `InitialClaimInflate`	**High correlation**

5 Sélection des données

Nous testerons des modèles de régressions plus tard. En cet effet, nous étudions la corrélation des variables numériques.



Il ressort du graphique ci-dessus que les variables `DaysWorkedPerWeek` et `HoursWorkedPerWeek` sont fortement corrélées, mais elle ne dépasse pas le seuil de 0.8 que nous nous fixons. Donc, nous n'avons pas de problème de multi-colinéarité dans nos données.

Il est d'usage en modélisation de la sinistralité de séparer les sinistres graves des sinistres attritionelles. Pour ce fait, nous définissons un seuil de sinistralité avec la méthode des Kmeans (technique d'apprentissage non supervisé utilisée pour la partition des données en k groupes ou clusters) afin de séparer les sinistres graves des sinistres attritionelles. Nous obtenons cette répartition ci-dessous.