Assur'Aimant

Objectif : Réalisez une étude exploratoire et un modèle pour prédire une prime d'assurance

Sommaire

- **Nettoyage des données**
- **Exploration du jeu de données**
- **Modélisation**
- **Démonstration du modèle avec un prototype d’application**

Nettoyage des données

Voici un échantillon des données.

```
data = pd.read_csv('donnee_brief.csv')
data.head()
```

0.7s — Python

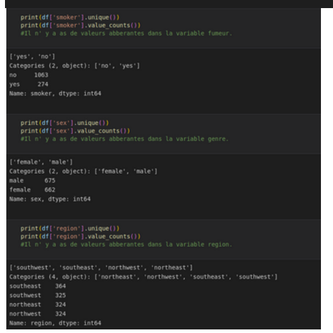
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Troisième étape : Vérification de valeur abberantes

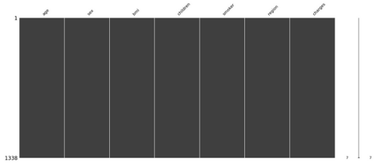
```
print(df['children'].unique())
print(df['children'].value_counts())
#Il n' y a as de valeurs abberantes dans la variable children
```

0.1s

```
[0 1 3 2 5 4]
0    573
1    324
2    240
3    157
4     25
5     18
Name: children, dtype: int64
```



Première étape : Vérifions s'il y a des valeurs manquantes



Après vérification , on a noté l'abscence de données manquantes. Voici un graphique missigno qui nous montrerai grâce à des lignes blanches horizontales s'il y avait des élements manquants.Comme il n' yen a pas , on confirme qu'il n'y en a pas.

Deuxième étape : Vérifier la présence d'élements en double.

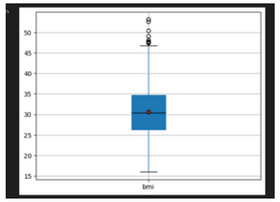
En explorant la base de données , on a noté la présence de deux élements identiques

```
data.loc[data.duplicated(keep=False),:]
#On recherche les element en double, et on vérifie si leur index
# est proche pour valider si cela est une erreur humaine
```

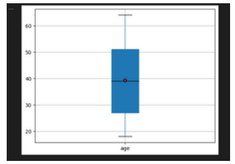
	age	sex	bmi	children	smoker	region	charges
195	19	male	30.59	0	no	northwest	1639.5631
581	19	male	30.59	0	no	northwest	1639.5631

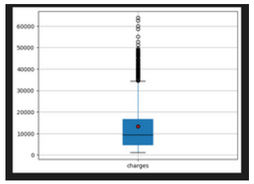
Les index ne sont pas proches , donc nous ne pouvons pas conclure si cela est une erreur de copie, il nous faudrait plus d'informations.
essayons de réflechir à la probabilités que deux personnes soit à ce point identiques.

La probabilité que 2 personnes aient le même indice de masse corporelle (IMC) est extrêmement faible. Étant donné que l'IMC est calculé à partir des valeurs de poids et de taille et que ces valeurs sont uniques pour chaque personne, la probabilité que 2 personnes aient le même IMC est très faible.
Nous considèrerons donc que cela est dû à une erreur et nous allons donc supprimer les doublons



L'age n'a pas de valeur atypique.





Les charges ont elles de nombreuses valeurs atypiques , ce qui semble logique , si l'on prend en compte les écarts brutaux de richesses aux états unis

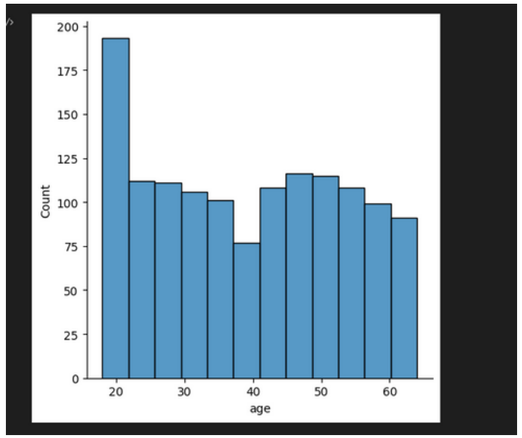
Pour la BMI , nous observons un certain nombre de valeurs aberrantes, ce qui semble correspondre avec les problèmes d'obésités aux étas Unis

Conclusion de l'exploration du jeu de données

Nous avons un jeu de donnée plûtot propre et organisé , mais qui va nécessité une exploration appronfondie , surtout en ce qui concerne les charges et la bmi , qui présente des données atypiques à étudiés.

Exploration des données

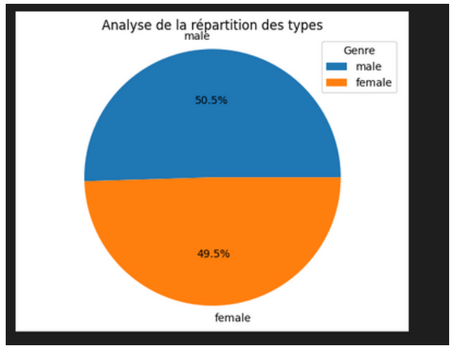
Analyse age



Cela ne ressemble pas à une distribution normale , mais plutot equiprobable , avec un pic pour les 18 ans , surement dû à une promotion pour la majorité, une pratique courante pour les banques .Il semble que la banque fasse des efforts pour chaque catégorie d'âge , afin d'attirer de manière égale les personnes de tout âge.

```
...   Moyenne
      39.20702541106129
      Mode
      0    18
      Name: age, dtype: int64
      Mediane
      39.0
      Etendue
      46
      Ecart-type
      14.049960379216154
      Coefficient de variation
      0.35835313268249896
```

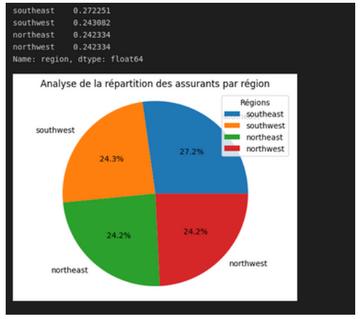
Analyse du genre




```
male      0.504862
female    0.495138
Name: sex, dtype: float64
```

Il y a une repartition relativement équiprobable , ce qui peut être dû à une politique égalitaire de l'assurance .

Analyse des régions.



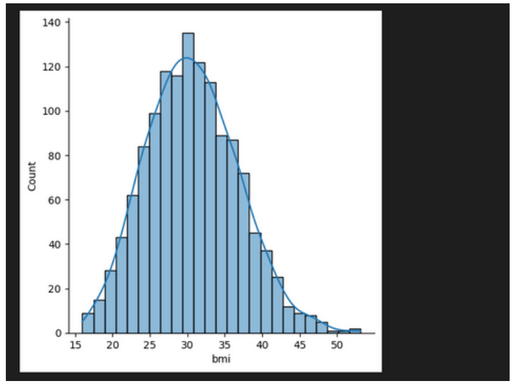
La répartions des populations venants des régions est plutôt équilibré, ce qui sous entend un politique égalitaire , hypothèse soutenue par la répartition homme femme, afin d'inclure des personnes des 4 régions.

Nous pouvons supposer une politique égalitaire pour notre assurance , et adapter nos choix pour adapter nos données pour nos prédictions.

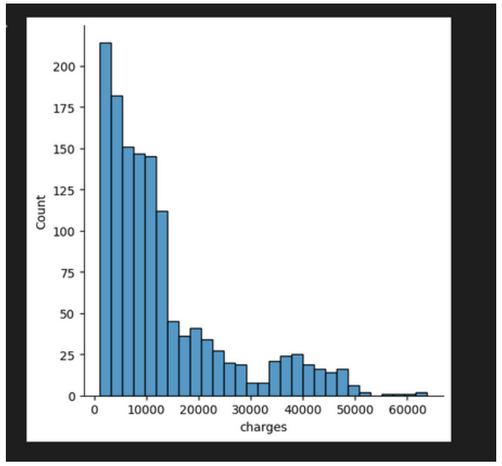
Analyse de la BMI

La BMI semble suivre une distribution normale, que l'on souligne ici via une approximation graphique.

```
Moyenne de la BMI
30.66345175766642
Mode de la BMI
0    32.3
Name: bmi, dtype: float64
Mediane de la BMI
30.4
Etendue de la BMI
37.17
Ecart-type de la BMI
6.100468409615801
Coefficient de variation de la BMI
0.198949174340445
```

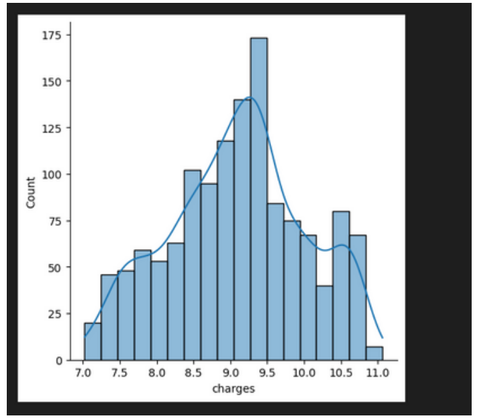


Analyse des charges



Les charges ne sembles pas suivre de distribution particulière

```
Moyenne des charges
13279.121486655948
Mode des charges
0        1121.87390
1        1131.50660
2        1135.94070
3        1136.39940
4        1137.01100
            ...
1332    55135.40209
1333    58571.07448
1334    60021.39897
1335    62592.87309
1336    63770.42801
Name: charges, Length: 1337, dtype: float64
Mediane des charges
9386.1613
Etendue des charges
62648.554110000005
Ecart-type des charges
12110.359656344175
Coefficient de variation des charges
0.9119850035647126
```

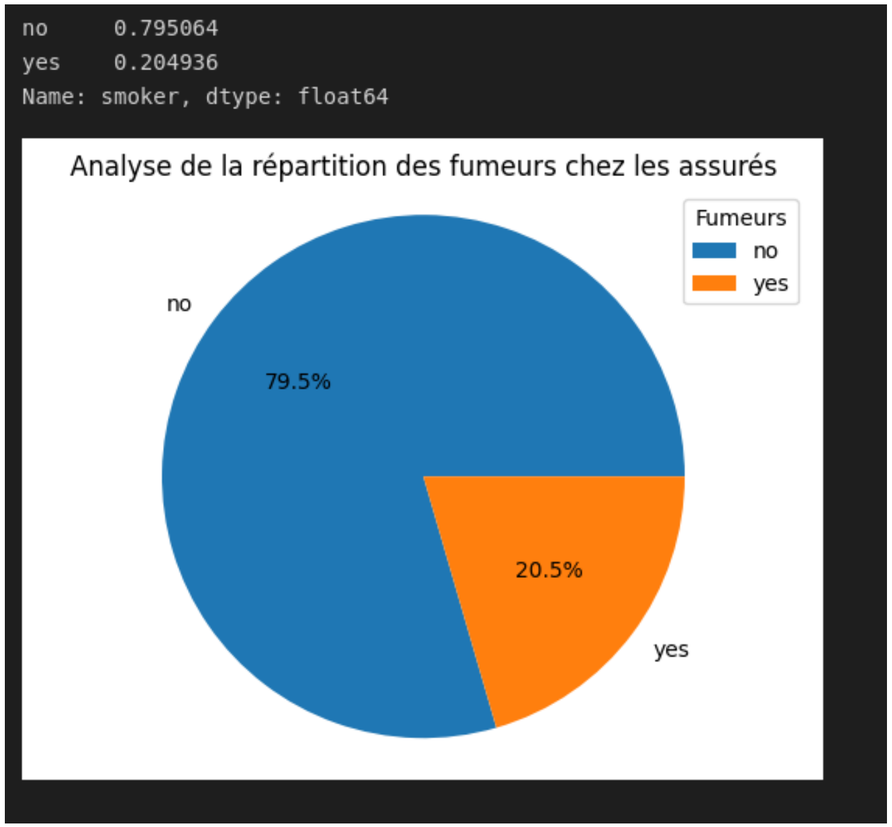


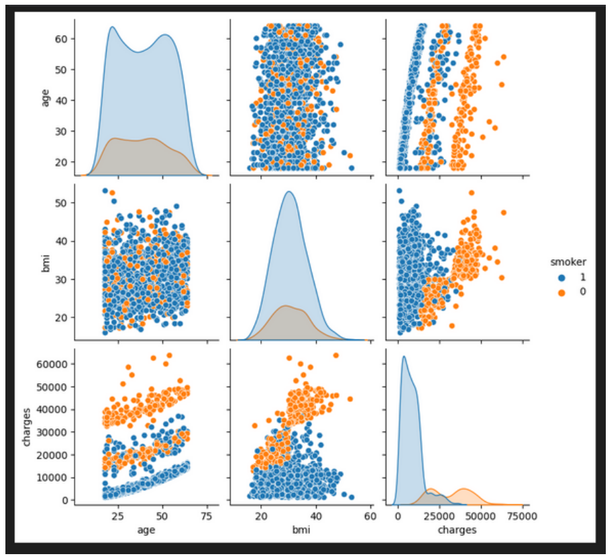
Cela m'a semblé bizarre , donc en effectuant une transformée logarithmique , on observe une ressemblance avec une loi normale .On peut considérer les valeurs comme ayant une trop grande disparité pour être facilement analysable, mais ayant quand même une relation derrière, une hypothèse soutenue par la disparité des richesses aux états-unis.Nous allons quand même prendre en compte ce phénomène dans notre analyse, en préferant un test kruskal , qui est plus "résistant" si la distribution n'est pas normale.

Analyse du fait de fumer.

La répartition des fumeurs est grandement inégale, ce qui peut sous entendre une politique restraignante pour les fumeurs. Cela peut être aussi lié à une proportion moindre des fumeurs dans la population.

```
no     0.795064
yes    0.204936
Name: smoker, dtype: float64
```

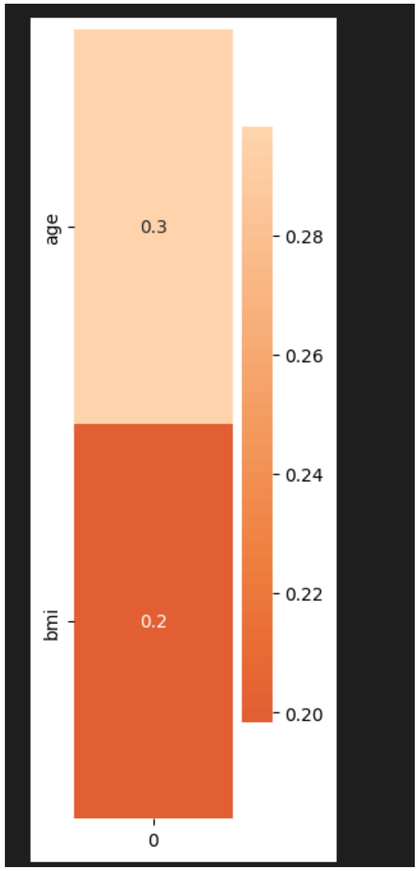




Comme le fait de fumer semble un être facteur aggravant, on va observer l'influence de fumer sur toutes nos variables.

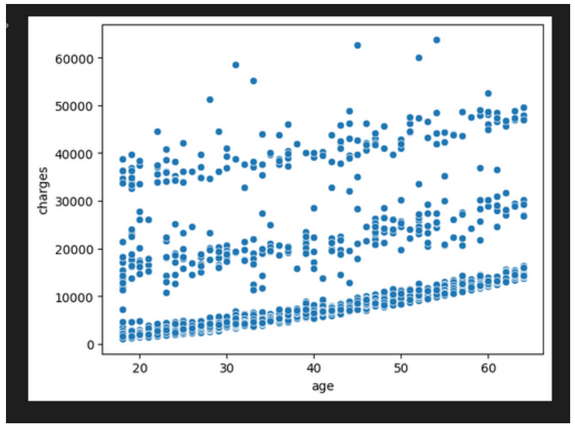
Analyse bivarié de la charge

Tout d'abord entre variable quantitatives.

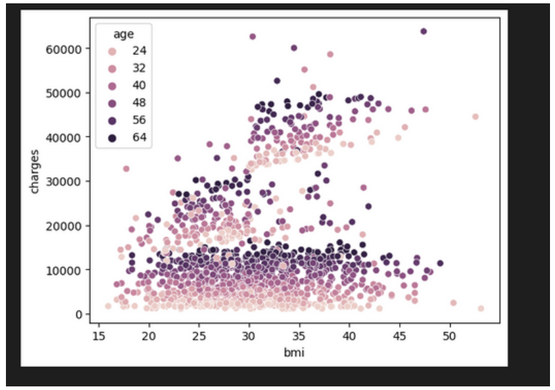



```
correlation charge et age: 0.2983082125097864
correlation charge et bmi: 0.19840083122624 94
```

Grâce à la corrélation, on remarque une relation modéré entre la charge et l'âge et la bmi



Cela est plus ou moins visible selon l'âge, ce qui peut induire que l'âge n'est pas un paramètre d'acceptation ou de rejet de prime mais on peut noter 3 zones , ce qui peut indiquer une distinction entre 3 catégorie d'âges, et donc d'une influence de l'âge sur les charges.Cela pourrait être une piste pour affiner notre modèle, en impliquant l'hypothèse que la santé influe sur la charge, selon certains facteurs à déterminer, notamment la catégorie d'âge



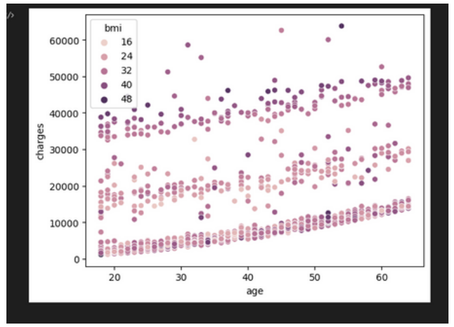
Entre bmi ,charge et age.

Lorsque l'on veut voir l'influence de la bmi sur l'âge , ce n'est pas aussi clair.La correlation age bmi est plus faible que la correlation charge bmi.Bien que la relation soit visible , elle peut être plus complexe et pas forcément symétrique.

```
correlation bmi et age: 0.1092718815485352
correlation bmi et charge: 0.1983409688336289
```

L'inverse n'est cependant pas vrai.Grâce à la densité des points , on peut déduire une influence notable de l'âge sur la charge en fonction de la BMI .Il y a un lien plus complexe lors de l'attribution des charges , dont les facteurs sont l'âge et la BMI , des indicateurs de santé.

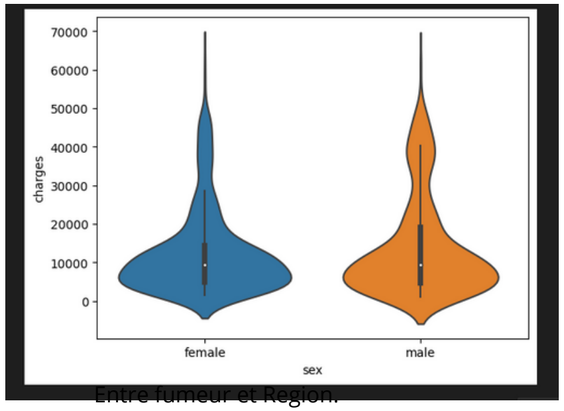
Cependant , cela ne veut pas dire que ces 2 variables influencent l'une l'autre de la même façon



Il peut être intéressant de changer la bmi selon des catégories pour affiner les tests.

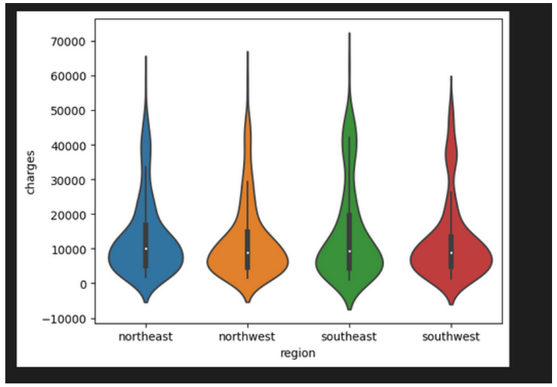
Entre catégorie.

Analyse genre /charge.



Il n'ya pas de différence entre homme et femmes .
De plus , nous obtenons les résultats suivants lorsque l'on effectue le test ANOVA pointbiserial:
pointbiserialr entre charge et genre
t-value: -0.05804449579031286
La faible valeur négative implique une très faible corrélation.On peut la considérer négligeable
p-value: 0.03382079199510838
La faible p -value assure la validité de notre test.

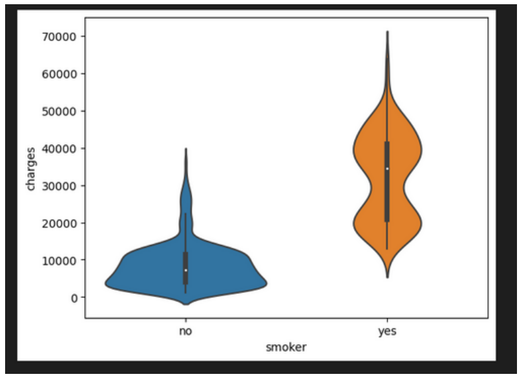
Entre fumeur et Region.



Ici , nous avons des graphiques similaires entre les regions.Cela nous indique que la région n'a pas d'influence sur l'attribution des primes.
Nous appliquons un test kruskal
T-test entre charge et region
t-value: 40.0894801326144
p-value: 1.4404664073876777e-275

Comme la p-value est inférieure à 0,05, notre test est significatif.
Le fait que la t-value est positif indique que les 2 régions ne sont par corrélés.On peut donc décider d'ignorer cet variable si nécessaire.

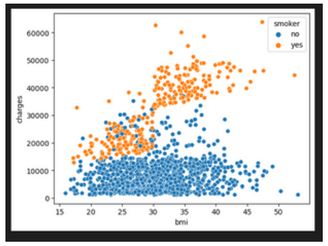
Entre fumeur et charge.

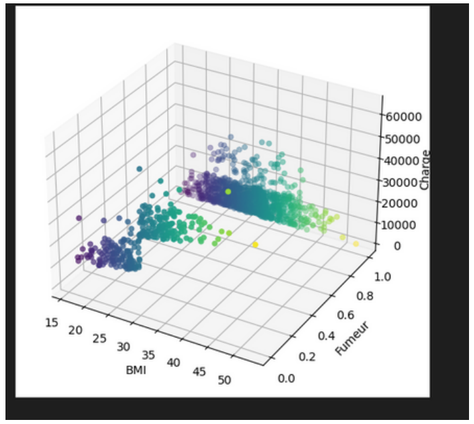


On remarque qu'il y a une forte relation entre fumeur et charge.Cela peut être dû à l'augmentation de la mortatlité à cause du fait de fumer , ou à une politique plus sévère pour les fumeurs (pour une politique de prévention)

On va comparer avec le test biserial qui est un test ANOVA entre variable quantitative et variable categorielle binaire
Le test nous renvoie :
mesure = 588.347
Nous avons donc une forte corrélation entre le fait de fumer et la charge
p_value = 0.000
La p value de 0 permet de confirmer la validité de notre test.

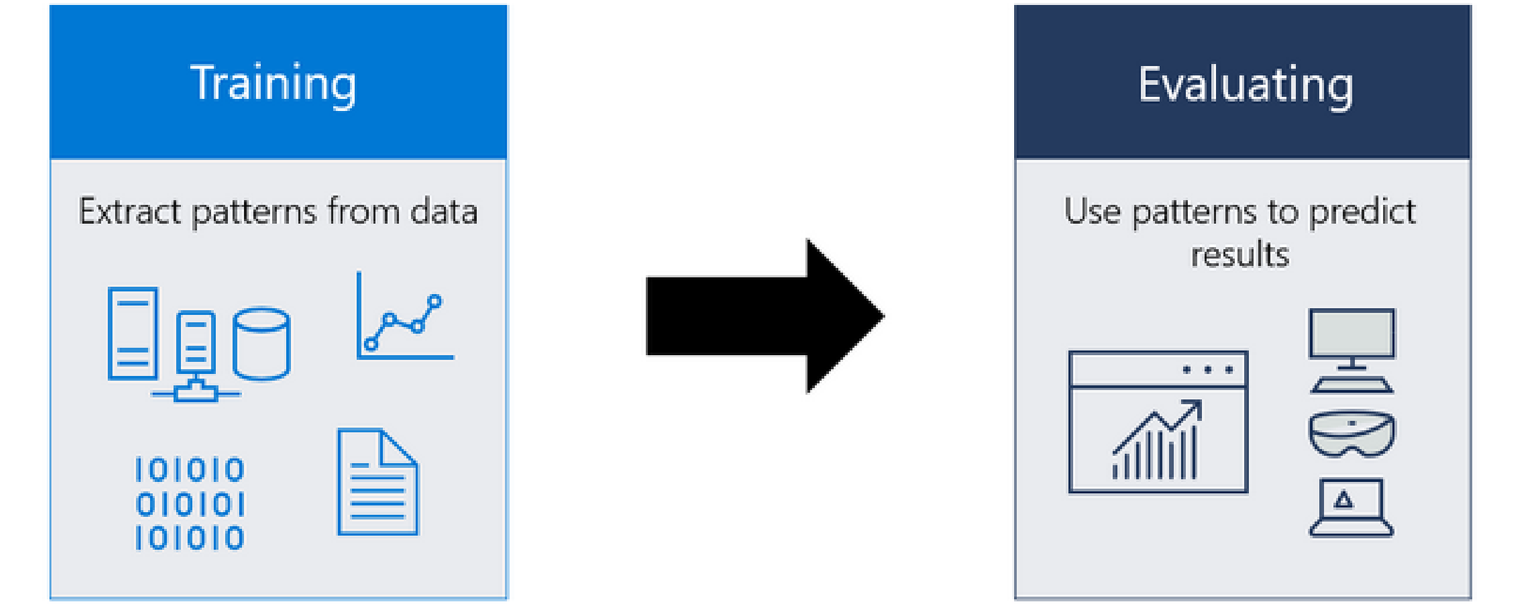
En étudiant plus attentivement les liens entre (charge,fumeur) et bmi et age , on remarque (surtout pour la bmi, charge et fumeur) une relation forte entre la BMI et le fait de fumer.On en déduit une politique de la banque concernant la distribution de charges en fonction de facteurs de santé aggravant, ce qui semble logique.





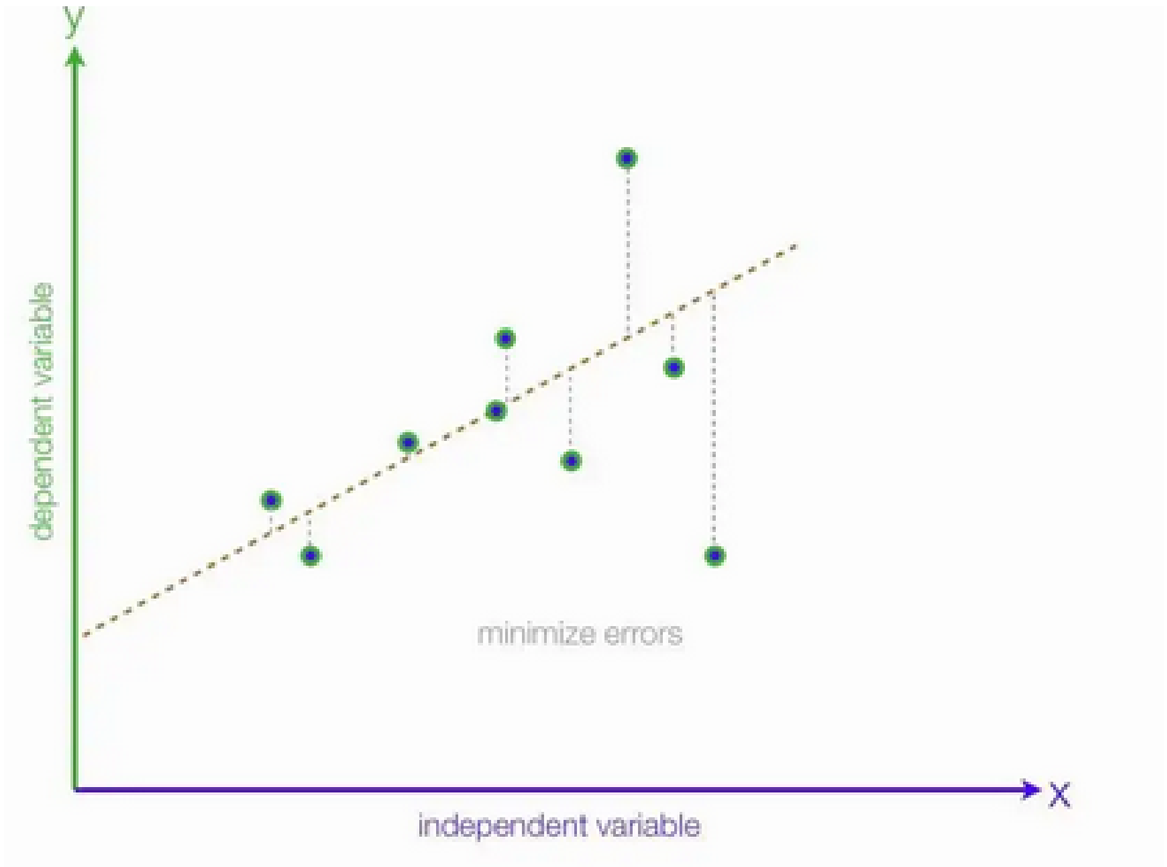
Entre bmi ,charge et fumeur.

Modélisation



Modèle de Machine Learning : Modèle entrainé sur un jeu de données, qui grâce à un algorithme mathématique et à l'entrainement et l'observation sur les données existantes, peut apprendre de celles-ci et prédire des données qu'il n'a encore jamais vues

La valeur a déterminer étant un nombre (prime d'assurance), il nous faut utiliser un modèle de Regression



Les deux modèles de Regression utilisés pour l'étude sont le modèle de Regression Linéaire et le modèle Elastic Net, qui lui permet de modifier les paramètres pour éviter le surapprentissage et donc permet de selectionner les variables les plus importantes pour la prédiction

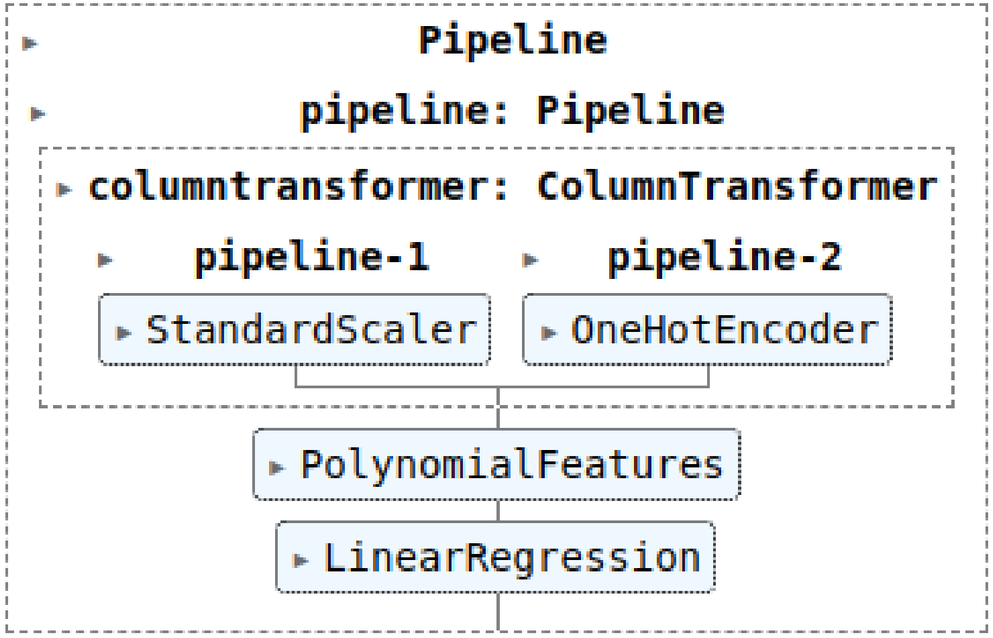
Grâce à nos observations pendant l'exploration du jeu de données, nous avons constatés que celui-ci était séparés en 2 types de données :

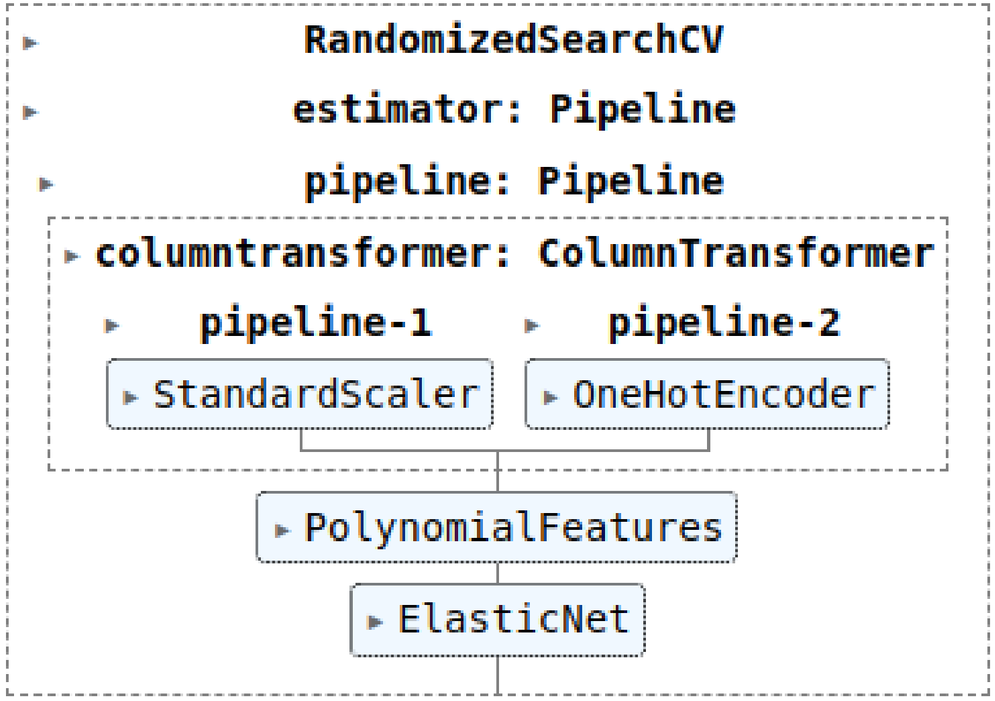
- **Des valeurs quantitatives (numériques) : age, imc et nombre d'enfants par client**
- **Des valeurs qualitatives (catégorielles) : sexe, statut fumeur et region**

Pour la réalisation de notre modèle :

- **Normaliser les données numériques afin d'adapter les valeurs a une échelle commune**
- **Encoder les valeurs catégorielles afin de les transformer en valeurs numériques et ainsi les rendre utilisable par notre modèle tout en conservant leurs informations**

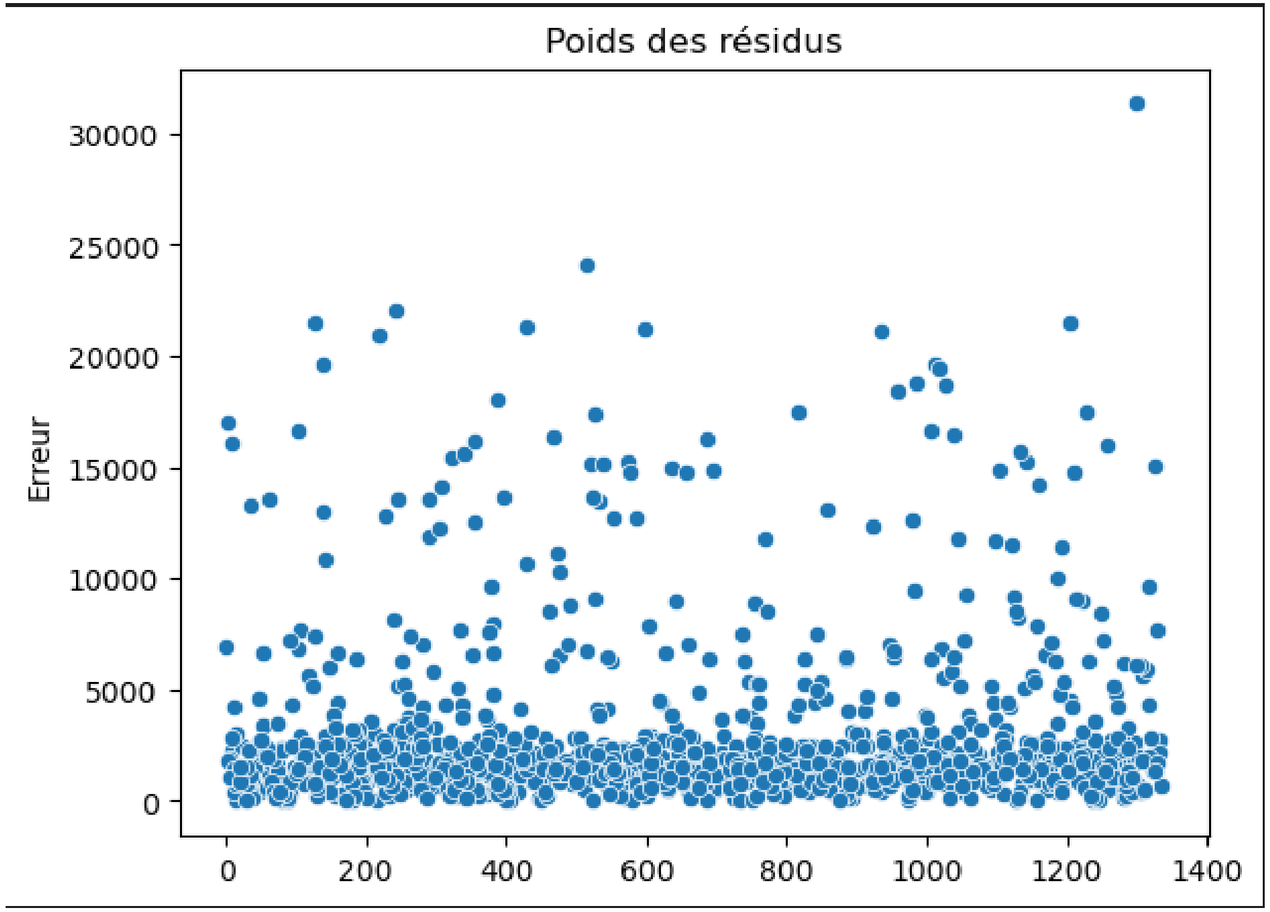
Pour effectuer cela,, nous avons utilliser un Pipeline qui est un objet qui permet de regrouper les étapes de transformation et de les appliquer sur nos deux types de catégories de données. Cela permet de faciliter la gestion des données et d'optimiser les performances de notre modèle.





Après entrainement de nos deux modèles, on obtient un score d'environ 82% de performance sur les deux. Pour la suite de notre étude nous avons décidé de continuer avec le modèle de Regression Linéaire celui-ci ayant un score légèrement meilleur

Afin améliorer le score de performance de notre modèle et éliminer la variabilité aléatoire, nous nous sommes penchés sur les résidus de celui-ci qui sont la différence entre les valeurs cibles rééles observées dans les données et les valeurs prévuees par le modèle



⚠ **Il est important de trouver un équilibre entre suppression majeure des résidus et la préservation de l'information contenue dans les données**

Après élimination d'une partie des résidus, on obtient un score final de 90% de performance sur notre modèle de Regression Linéaire