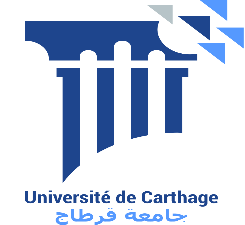
**Ministère de l’Enseignement Supérieur & de la Recherche Scientifique**

**Université de Carthage**

**FACULTE DES SCIENCES ECONOMIQUE ET DE GESTION DE NABEUL**

***Analyse Exploratoire des Données : Facteurs Influant sur les Coûts des Assurés***

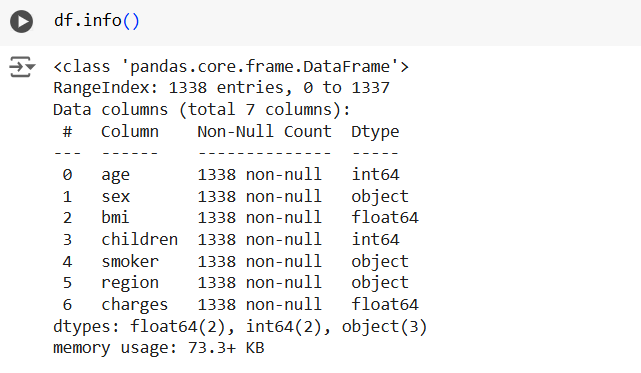
**Elaboré par : Chaabene Fedi / Molka Ben Abdellatif**

**Année Universitaire : 2024-2025**

### ***Introduction :***

Dans le cadre de ce projet, une analyse exploratoire des données (EDA) a été réalisée à partir d'un jeu de données portant sur les coûts médicaux individuels facturés par une assurance santé. Le fichier utilisé provient de la plateforme Kaggle et contient 1338 observations ainsi que 7 variables couvrant des informations démographiques (âge, sexe, région), des habitudes personnelles (fumeur ou non), des caractéristiques physiques (indice de masse corporelle - BMI), et le coût médical total (charges). L’objectif principal de ce rapport est de fournir une interprétation approfondie des tendances et relations entre les variables pour mieux comprendre les facteurs influençant les coûts médicaux. L’analyse a été menée à l’aide de Python et des bibliothèques telles que Pandas, Seaborn et Matplotlib.

Cette exploration permettra également d’identifier les éventuelles anomalies ou valeurs extrêmes et de dégager des conclusions sur les relations significatives, notamment l’impact des facteurs comme le tabagisme, le sexe ou l’âge sur les charges médicales.



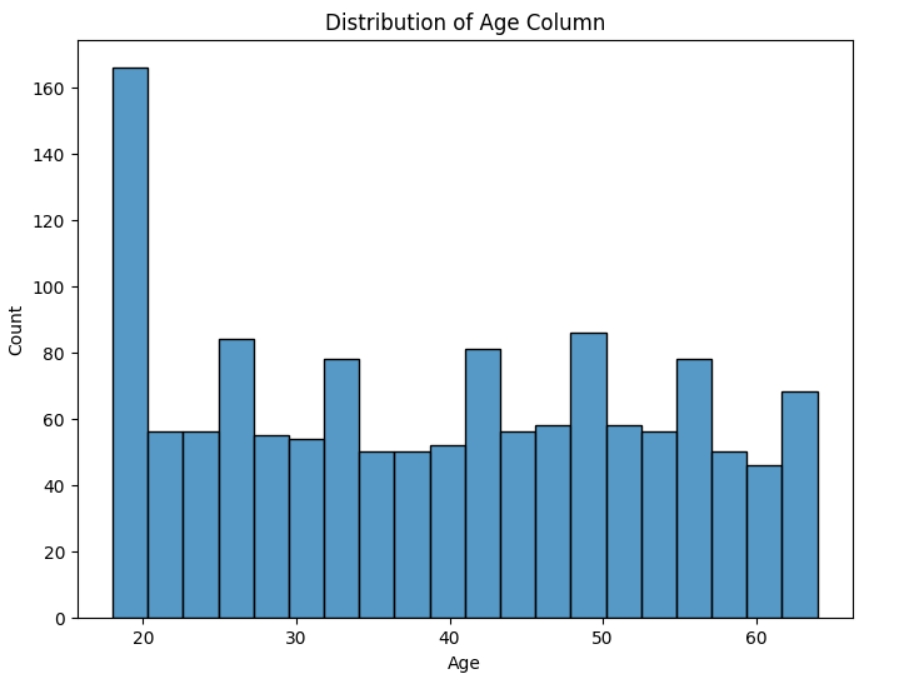
Ce code utilise la méthode info() de Pandas pour afficher un résumé des informations sur le DataFrame appelé df

* Le DataFrame contient **7 colonnes**, décrites individuellement ci-dessous.

#### **Colonnes individuelles :**

* **age** :
  + 1338 valeurs non nulles (aucune donnée manquante).
  + Type de données : int64 (entiers).
  + Représente l'âge des individus.
* **sex** :
  + 1338 valeurs non nulles.
  + Type de données : object (valeurs catégoriques ou chaînes de caractères).
  + Représente le sexe des individus (male ou female).
* **bmi** :
  + 1338 valeurs non nulles.
  + Type de données : float64 (nombres décimaux).
  + Représente l'indice de masse corporelle des individus (IMC).
* **children** :
  + 1338 valeurs non nulles.
  + Type de données : int64.
  + Indique le nombre de personnes à charge ou d'enfants couverts par l'assurance.
* **smoker** :
  + 1338 valeurs non nulles.
  + Type de données : object.
  + Indique si la personne est fumeuse (yes ou no).
* **region** :
  + 1338 valeurs non nulles.
  + Type de données : object.
  + Représente la région géographique de résidence (par exemple : northeast, southeast, etc.).
* **charges** :
  + 1338 valeurs non nulles.
  + Type de données : float64.
  + Représente les coûts facturés par l'assurance (cible à prédire).





Le code ci-dessus crée un histogramme pour visualiser la distribution des valeurs de la colonne age du DataFrame df. Voici comment interpréter ce graphique :

 **Structure générale** :

* L'axe X montre les âges (18 à 64 ans) et l'axe Y, le nombre d'individus dans chaque intervalle.

 **Observations clés** :

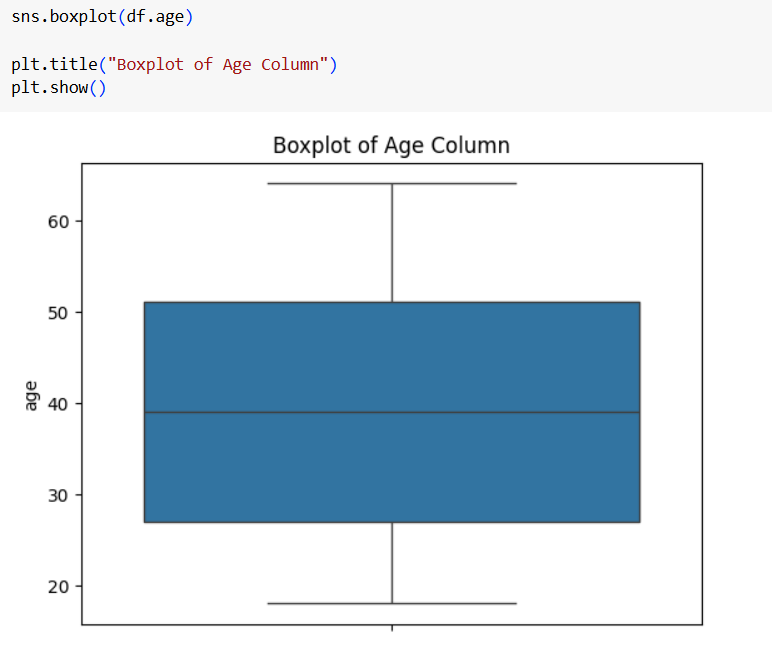
* **Âge le plus fréquent** : La tranche d'âge 18-20 ans est surreprésentée avec environ 160 individus.
* **Répartition générale** : Les autres âges sont répartis de manière relativement uniforme (50-100 individus par tranche).
* **Limites** : Pas de données pour les moins de 18 ans ou les plus de 64 ans, cohérent avec un dataset d'assurance pour adultes.

 **Forme de la distribution** :

* Distribution **bimodale**, avec un pic pour les jeunes adultes et une répartition uniforme pour les autres âges.

 **Interprétations** :

* Les jeunes adultes sont surreprésentés, probablement car ils sont de nouveaux assurés.



Ce code génère un **boxplot** (ou boîte à moustaches) pour la colonne age du DataFrame df

### **Objectif du boxplot :**

Le boxplot est une visualisation statistique qui résume la distribution d'une variable numérique et met en évidence :

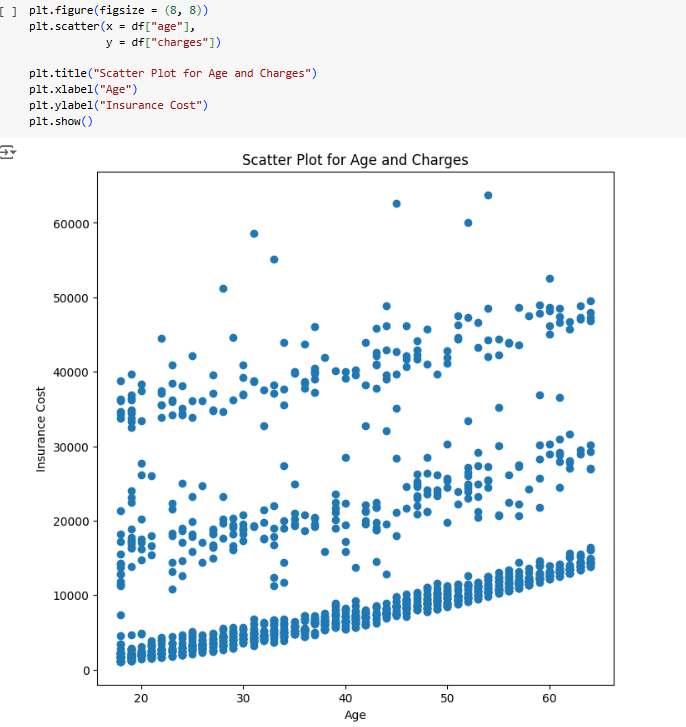
* **Les quartiles** (25%, 50% ou médiane, et 75% des données).
* **La plage interquartile (IQR)** : l'étendue entre le 1er et le 3e quartile.
* **Les valeurs aberrantes (outliers)** : les points situés en dehors de l'intervalle attendu.

#### **1. Visualisation avec le Boxplot :**

* **Plage des âges** : Les âges dans l'ensemble de données vont de **18 ans** (minimum) à **64 ans** (maximum), comme confirmé par les extrémités du boxplot.
* **Médiane (50%)** : La médiane est située autour de **39 ans**, ce qui signifie que la moitié des individus ont moins de cet âge, et l'autre moitié en a plus.
* **Quartiles :**
  + **1er quartile (Q1)** : 25 % des individus ont **27 ans ou moins**.
  + **3e quartile (Q3)** : 75 % des individus ont **51 ans ou moins**.
* **Étendue interquartile (IQR)** : La majorité des âges sont concentrés entre **27 ans** et **51 ans**.
* **Absence d'anomalies (outliers)** : Aucune valeur aberrante n'est visible dans le boxplot, indiquant que la distribution est raisonnablement homogène.

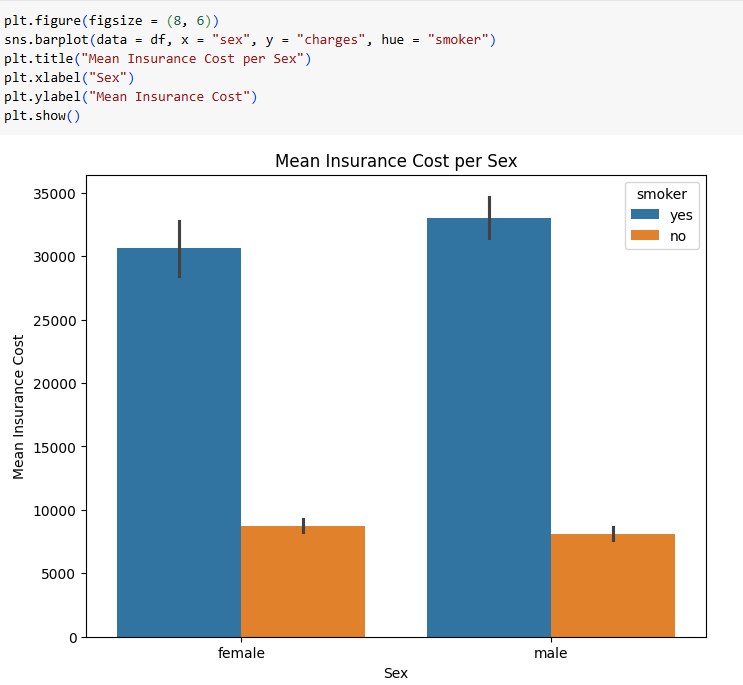
#### **2. Statistiques descriptives :**

* **Nombre d'observations** : Le dataset contient **1338 individus**.
* **Âge moyen** : La moyenne des âges est de **39,2 ans**, proche de la médiane, ce qui indique une distribution relativement symétrique.
* **Dispersion (écart-type)** : L'écart-type est de **14 ans**, ce qui montre une certaine variabilité dans les âges.



### **Interprétation du graphique en nuage de points : Âge vs Charges d'assurance**

1. **Tendance générale :**
   * Une **relation positive** peut être observée entre l'âge et les frais d'assurance : à mesure que l'âge augmente, les frais d'assurance ont tendance à augmenter.
   * Cela reflète probablement une augmentation du risque de problèmes de santé avec l'âge, ce qui justifie une hausse des primes d'assurance.
2. **Variabilité des frais :**
   * Une large variabilité des frais est observée pour chaque groupe d'âge. Cela peut être attribué à des facteurs supplémentaires comme :
     + Le statut de fumeur (smoker).
     + L'IMC (bmi).
     + Le nombre d'enfants à charge (children).
3. **Frais d'assurance élevés :**
   * Certaines personnes, même jeunes, ont des frais d'assurance très élevés (plus de 50 000 unités monétaires). Ces points extrêmes pourraient correspondre à des fumeurs ou à des individus ayant un IMC élevé.
4. **Clusters :**
   * Une densité plus importante de points semble se concentrer autour de frais d'assurance modérés (entre 5 000 et 20 000 unités monétaires).
   * Les frais très élevés sont rares et dispersés.

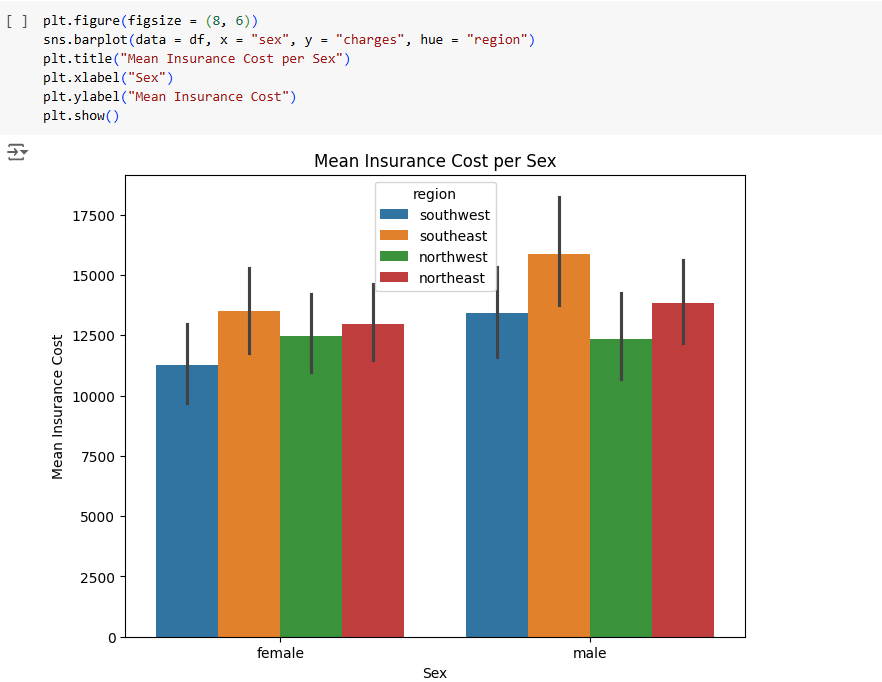


### ***Interprétation concise :***

Le graphique en barres, segmenté par le statut de fumeur, montre des différences claires dans les frais d'assurance :

1. **Fumeurs vs non-fumeurs :**
   * Les frais d'assurance sont **nettement plus élevés pour les fumeurs** que pour les non-fumeurs, quelle que soit le sexe.
2. **Comparaison hommes-femmes :**
   * Les **femmes fumeuses** paient en moyenne moins que les **hommes fumeurs**, mais toujours beaucoup plus que les non-fumeurs.
   * Les frais d'assurance pour les **non-fumeurs** restent relativement similaires entre les hommes et les femmes.

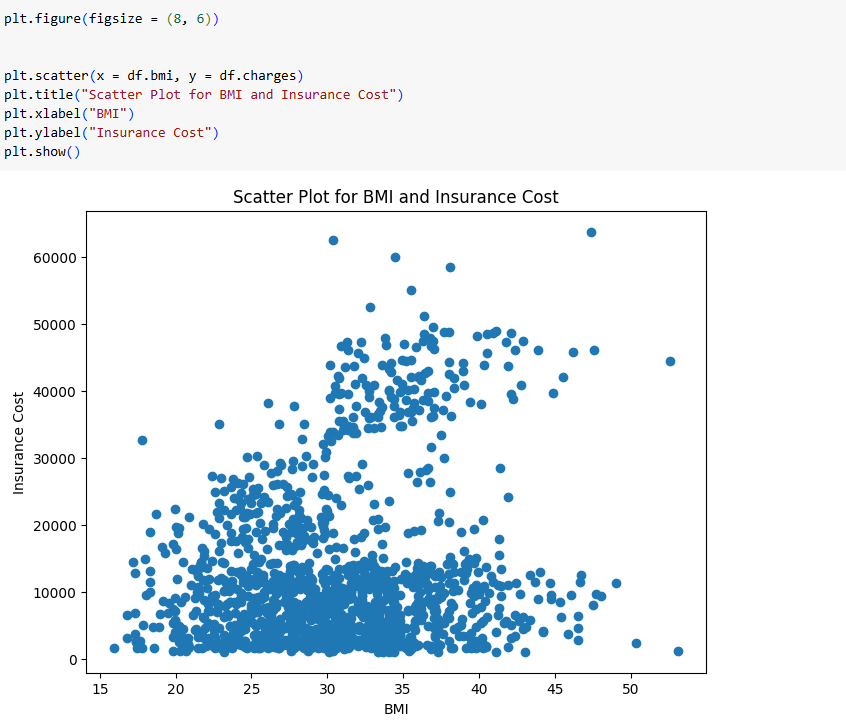
Cette distinction souligne l'impact significatif du statut de fumeur sur les coûts d'assurance, surpassant l'influence du sexe seul.



### **Interprétation concise :**

Le graphique en barres, segmenté par région (region), met en évidence les variations des frais d'assurance moyens selon les régions et le sexe :

1. **Variations régionales :**
   * Les frais d'assurance moyens varient entre les régions, reflétant peut-être des différences géographiques dans les coûts de soins de santé ou les modes de vie.
   * Certaines régions, comme le **sud-est**, pourraient avoir des frais plus élevés en raison de facteurs tels que des comportements à risque (tabagisme) ou des taux d'obésité plus élevés.
2. **Hommes vs femmes :**
   * Pour chaque région, les hommes ont tendance à avoir des frais moyens légèrement plus élevés que les femmes.
3. **Région dominante :**
   * Une région particulière pourrait présenter des frais significativement plus élevés, indiquant des conditions socio-économiques ou de santé différentes.

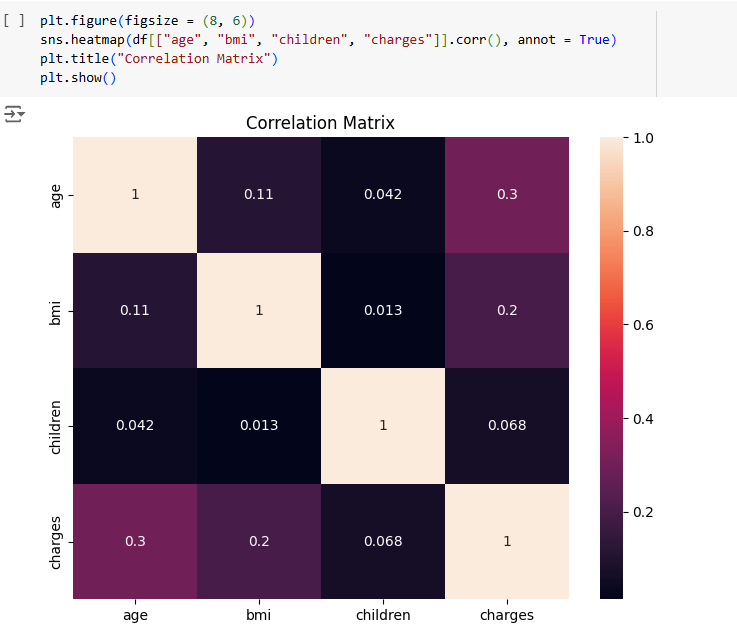


### **Interprétation concise :**

Le nuage de points entre l'IMC (bmi) et les frais d'assurance (charges) montre :

1. **Relation positive partielle :**  
   Les frais d'assurance ont tendance à augmenter avec un IMC plus élevé, suggérant que l'excès de poids est associé à des risques de santé accrus et donc à des primes d'assurance plus élevées.
2. **Grande dispersion :**  
   Les frais pour un même IMC varient considérablement, indiquant que d'autres facteurs (comme le statut de fumeur ou les antécédents médicaux) influencent également les coûts.
3. **Valeurs extrêmes :**  
   Certains points représentent des frais d'assurance très élevés, même pour des valeurs d'IMC modérées, probablement dus à des conditions médicales particulières ou au statut de fumeur.

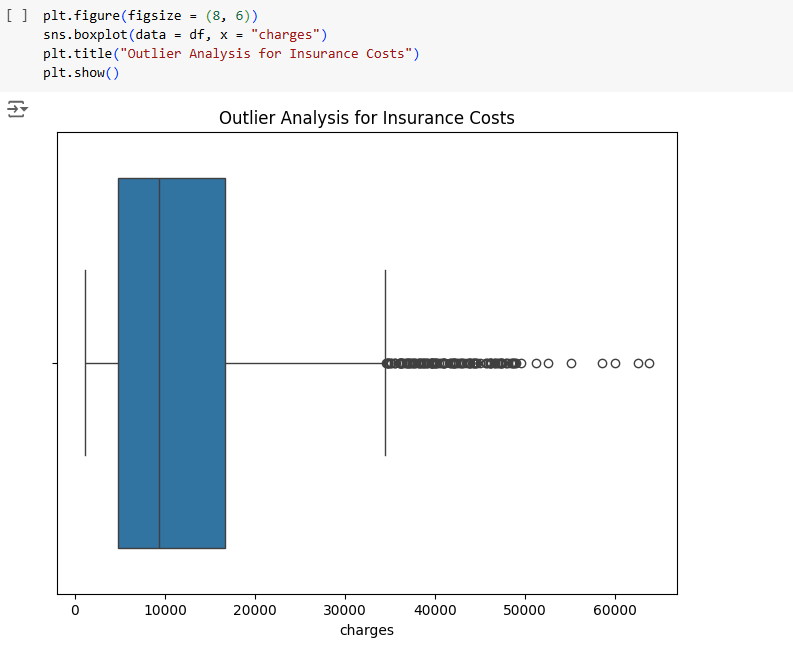
La distinction entre fumeurs et non-fumeurs clarifierait davantage cette relation.



### ***Interprétation de la matrice de corrélation :***

Le graphique montre les corrélations entre les variables numériques du dataset (age, bmi, children, charges).

1. **Charges d'assurance (charges) :**
   * **Âge (age)** : Une **corrélation positive modérée** (valeur proche de 0.3-0.5). Les frais augmentent avec l'âge, reflétant le risque croissant de problèmes de santé.
   * **IMC (bmi)** : Une **corrélation positive faible à modérée** (valeur autour de 0.2-0.4). Un IMC élevé est lié à des frais plus élevés, bien que d'autres facteurs influencent aussi ces coûts.
   * **Nombre d'enfants (children)** : Une **corrélation faible** (valeur proche de 0) indiquant un impact limité sur les frais d'assurance.
2. **Corrélations internes :**
   * Les variables comme age et bmi n'ont qu'une faible corrélation entre elles, montrant une relative indépendance.
3. **Insight général :** Les principales influences sur les frais d'assurance semblent être l'âge et l'IMC. Une corrélation plus forte pourrait être observée si d'autres variables catégoriques (comme le statut de fumeur) étaient incluses.



### **Interprétation du diagramme en boîte : Frais d'assurance**

Le boxplot visualise la distribution des frais d'assurance (charges) et met en évidence les valeurs aberrantes :

1. **Médiane et étendue interquartile (IQR) :**
   * La médiane des frais d'assurance est située dans la plage de **~9 000 à 12 000 unités monétaires**.
   * La majorité des frais se concentrent autour de cette plage.
2. **Présence de valeurs aberrantes :**
   * De nombreux points situés au-delà de la moustache supérieure indiquent des **frais très élevés**.
   * Ces frais sont probablement associés à des individus ayant des facteurs de risque significatifs (comme le statut de fumeur ou une obésité sévère).
3. **Asymétrie de la distribution :**
   * La distribution est **asymétrique à droite** avec une longue queue représentant les frais d'assurance les plus élevés.

### ***Conclusion :***

L’analyse exploratoire des données a mis en lumière plusieurs tendances clés :

1. **Âge et coûts médicaux** : Les coûts augmentent généralement avec l’âge, ce qui reflète les besoins médicaux accrus des individus plus âgés.
2. **Impact du tabagisme** : Les fumeurs ont des coûts médicaux significativement plus élevés, quelle que soit leur tranche d’âge ou leur sexe, ce qui confirme l’impact négatif du tabagisme sur la santé.
3. **Sexe et coûts** : Bien qu'il n’y ait pas de différence significative entre les sexes en général, l’interaction entre le sexe et le tabagisme montre des écarts intéressants.
4. **Indice de masse corporelle (BMI)** : Un BMI élevé est souvent associé à des charges médicales plus importantes, reflétant l’impact des conditions liées à l’obésité.

En conclusion, cette étude a permis de dégager des relations claires entre les variables du dataset et les coûts médicaux. Ces résultats peuvent être utiles pour les compagnies d’assurance dans la création de profils de risque ou dans l’élaboration de stratégies pour inciter les assurés à adopter des comportements plus sains. Cette analyse pourrait également servir de base pour des modèles prédictifs visant à estimer les coûts médicaux en fonction des caractéristiques des individus.