المدرسة العليا للتكنولوجيا -أكادير
ⵜⵉⵎⵙⵙⵓⵔⵉⵜ ⵜⴰⵙⵙⴷⴰⵡⵜ ⵏ ⵜⵙⵙⵏⵜⵉⵔⵜ - ⴰⴳⴰⴷⵉⵔ
ÉCOLE SUPÉRIEURE DE TECHNOLOGIE - AGADIR

RAPPORT : le tabagisme et le cancer du poumon

Nom : **SAADIA BOUTIRGUINE**

Date : 21 Octobre 2025

DATASET :Les relations entre le tabagisme et le cancer du poumon

Objectifs de l'Analyse

L'objectif principal de cette analyse est de mieux comprendre les relations entre le tabagisme et le cancer du poumon à travers un dataset de 4000 individus. Plus précisément, nous voulons :

- Identifier les variables quantitatives et qualitatives importantes.
- Étudier les corrélations entre le nombre de cigarettes consommées, les années de tabagisme et le risque de cancer.
- Visualiser la distribution de l'âge, du BMI et d'autres facteurs de risque.
- Comprendre l'effet du tabagisme passif et de la consommation d'alcool sur la santé.
- Fournir des conclusions basées sur les données pour appuyer la prévention et la sensibilisation.

SOMMAIRE

1. Introduction

La visualisation des données joue un rôle fondamental dans le domaine de l'analyse des données modernes. À mesure que la quantité d'informations produites augmente de façon exponentielle, il devient essentiel de pouvoir interpréter rapidement et efficacement ces données. La visualisation permet de transformer des chiffres bruts et complexes en représentations graphiques claires et intuitives, rendant les informations compréhensibles pour un large public, qu'il s'agisse de décideurs, de chercheurs ou de professionnels du domaine. E Dans ce contexte, des outils comme Matplotlib, Seaborn, ou Power BI offrent la possibilité de créer une grande variété de visualisations adaptées aux besoins de l'analyse : histogrammes, boxplots, nuages de points, heatmaps, et bien d'autres encore. Ces outils ne servent pas uniquement à illustrer les données, mais également à raconter une histoire à travers elles, à mettre en évidence des patterns et à rendre l'information plus accessible et engageante. La visualisation devient ainsi un langage universel entre les analystes et les décideurs, un pont qui relie la complexité des données à la simplicité des insights exploitables.

Le tabagisme, en particulier, représente un problème de santé publique majeur. Il est reconnu comme l'une des principales causes de maladies chroniques et de cancers, notamment le cancer du poumon. Les conséquences du tabagisme ne se limitent pas aux fumeurs actifs : l'exposition au tabagisme passif constitue également un facteur de risque significatif. Comprendre les liens entre le tabagisme et le développement du cancer est essentiel pour la prévention, la sensibilisation et la mise en place de politiques de santé efficaces. C'est dans ce cadre que s'inscrit notre analyse, visant à étudier les interactions entre le tabagisme et d'autres facteurs de risque tels que l'âge, le sexe, l'IMC, la consommation d'alcool ou le niveau d'activité physique.

Cette étude repose sur un dataset de 4000 individus, soigneusement nettoyé et prêt pour l'analyse. Il comprend des informations détaillées sur chaque participant, telles que l'âge, le sexe, le statut de fumeur, le nombre de cigarettes consommées quotidiennement, les années de tabagisme, l'exposition au tabagisme passif, la consommation d'alcool, l'IMC, le niveau d'activité physique et le diagnostic de cancer du poumon. L'objectif principal de cette analyse est double : d'une part, maîtriser les techniques de visualisation univariée et bivariée pour explorer les données de manière approfondie, et d'autre part, identifier et comprendre les facteurs de risque associés au tabagisme et à la survenue du cancer.

Grâce à cette approche, il devient possible de mettre en lumière des relations et tendances qui pourraient passer inaperçues dans une analyse purement statistique. Les visualisations permettent d'illustrer les distributions, les corrélations et les anomalies, tout en offrant un support visuel puissant pour la communication des résultats. L'étude vise également à fournir des recommandations basées sur les données pour améliorer la sensibilisation et la prévention du tabagisme, démontrant ainsi l'importance de la visualisation comme outil d'analyse et de décision dans le domaine de la santé publique.

PARTIE 1 : Chargement et Exploration Basique

L'objectif de cette partie est de charger le jeu de données, d'examiner sa structure et d'effectuer une première exploration descriptive. Cette étape est essentielle pour comprendre le contenu du dataset, identifier les valeurs manquantes et analyser les types de variables disponibles.

Présentation du Jeu de Données

Le dataset utilisé dans cette étude contient 4000 individus et 12 colonnes principales :

AGE : âge des participants. **GENDER** : sexe (Male/Female). **SMOKING_STATUS** : statut de fumeur (Fumeur, Non-fumeur, Ancien fumeur). **CIGARETTES_PER_DAY** : nombre de cigarettes consommées par jour. **YEARS_SMOKING** : nombre d'années de tabagisme. **EXPOSURE_TO_SECONDHAND_SMOKE** : exposition au tabagisme passif. **ALCOHOL_CONSUMPTION** : consommation d'alcool. **BMI** : indice de masse corporelle. **PHYSICAL_ACTIVITY_LEVEL** : niveau d'activité physique. **LUNG_CANCER** : diagnostic de cancer du poumon (Oui/Non). **SURVIVAL_YEARS** : années de survie après diagnostic. **COMMENTAIRES** : observations supplémentaires (si présentes).

Cette description permet d'avoir une idée claire des types de variables disponibles et de préparer les étapes suivantes de l'analyse exploratoire.

Outils et Librairies Utilisés

Pour réaliser cette analyse de données, j'ai utilisé plusieurs outils et librairies puissantes sur **Jupyter Notebook**, permettant de manipuler et visualiser les données de manière efficace et moderne :

- "🐍 Python : langage principal pour l'analyse et la manipulation des données."

  "📊 Pandas : gestion et traitement des DataFrames."

  "📈 Matplotlib : création de graphiques statiques et personnalisés."

  "🎨 Seaborn : visualisation statistique avancée et attractive."

  "🔢 Numpy & Scipy : calcul scientifique et traitement numérique."

Téléchargement du Dataset

Pour commencer l'analyse, le dataset a été téléchargé depuis une source fiable et stocké localement pour être utilisé dans **Jupyter Notebook**:

Dataset Complet :

	Patient_ID	AGE	GENDER	SMOKING_STATUS	CIGARETTES_PER_DAY	YEARS_SMOKING	EXPOSURE_TO_SECONDHAND_SMOKE	ALCOHOL_CONSUMPTION	BMI
0	1	63	M	Former	28	38	No	Yes	32.2
1	2	76	M	Current	30	3	No	No	25.4
2	3	53	F	Former	24	34	No	Yes	18.1
3	4	39	F	Current	23	46	No	Yes	32.9
4	5	67	F	Never	18	35	No	No	32.9
...	...	...	...	...	...	...	...	...	...
3995	3996	54	F	Former	38	5	Yes	Yes	18.3
3996	3997	57	M	Former	9	45	No	No	30.4
3997	3998	60	M	Current	16	32	No	Yes	30.1
3998	3999	42	M	Former	6	48	No	No	29.1
3999	4000	61	F	Never	39	16	No	Yes	32.9

PHYSICAL_ACTIVITY_LEVEL	LUNG_CANCER	SURVIVAL_YEARS
Moderate	Yes	2.5
High	No	9.9
Low	Yes	6.7
High	Yes	2.9
Moderate	No	6.8
...	...	...
Low	No	2.6
Low	No	4.0
High	No	8.4
Low	No	6.3
High	Yes	1.7

Prendre des informations générales sur mon dataset :

Rapelle

“Smoking & Cancer” est centré sur l’étude des liens entre le tabagisme et le développement du cancer, plus précisément le cancer du poumon.

```python
# Dimensions du dataset
print("Shape:", df.shape)


print('='*70)
# Vérification des valeurs manquantes
print(df.isnull().sum())
print('='*70)
print(df.columns)
print('='*70)
print(df.info())
print('='*70)
```

output :

Le dataset contient 4000 individus et 12 colonnes principales. Toutes les colonnes sont complètes, sans valeurs manquantes. Les types de données sont variés : 4 colonnes numériques entières, 2 colonnes numériques flottantes et 6 colonnes de type chaîne (texte).

Colonnes et types de données :

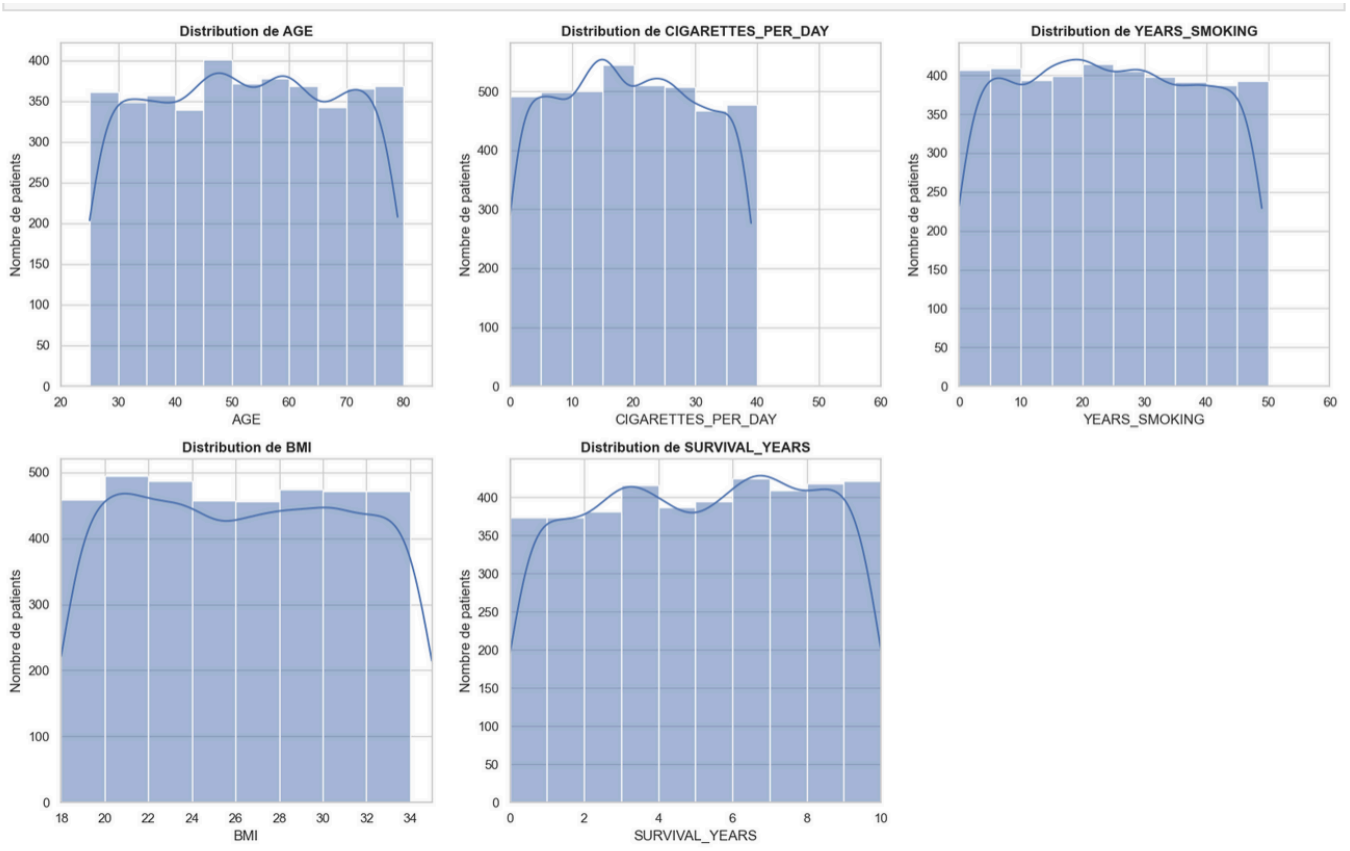
- Patient_ID : int64
- AGE : int64
- GENDER : object
- SMOKING_STATUS : object
- CIGARETTES_PER_DAY : int64
- YEARS_SMOKING : int64
- EXPOSURE_TO_SECONDHAND_SMOKE : object
- ALCOHOL_CONSUMPTION : object
- BMI : float64
- PHYSICAL_ACTIVITY_LEVEL : object
- LUNG_CANCER : object
- SURVIVAL_YEARS : float64

Cette première exploration montre que le dataset est propre et prêt à être analysé, avec toutes les variables correctement renseignées et des types de données adaptés à une analyse statistique et visuelle.

	Patient_ID	AGE	CIGARETTES_PER_DAY	YEARS_SMOKING	BMI	SURVIVAL_YEARS
count	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000
mean	2000.500000	52.139000	19.292500	24.346500	26.414500	5.073175
std	1154.844867	15.799291	11.346745	14.380565	4.909527	2.876802
min	1.000000	25.000000	0.000000	0.000000	18.000000	0.000000
25%	1000.750000	39.000000	10.000000	12.000000	22.100000	2.600000
50%	2000.500000	52.000000	19.000000	24.000000	26.400000	5.100000
75%	3000.250000	66.000000	29.000000	37.000000	30.625000	7.500000
max	4000.000000	79.000000	39.000000	49.000000	35.000000	10.000000

DISTRIBUTIONS DES VARIABLES NUMÉRIQUES :

Pour mieux comprendre le comportement des patients dans le dataset, nous nous concentrons sur trois variables numériques principales :('AGE', 'CIGARETTES_PER_DAY', 'YEARS_SMOKING', 'BMI', 'SURVIVAL_YEARS')



1️⃣ AGE

L'histogramme montre la répartition des âges dans la population étudiée.

La population étudiée est majoritairement d'âge moyen à avancé, ce qui est cohérent avec une étude sur des maladies chroniques ou cardiovasculaires qui augmentent avec l'âge.

Étendue : environ 20 à 85 ans

Distribution relativement uniforme avec une légère concentration autour de 50 ans

Pic visible autour de 50-55 ans avec environ 400 patients

2️⃣ CIGARETTES_PER_DAY

Montre combien de cigarettes par jour les patients fument.

Distribution asymétrique avec un pic important entre 15-25 cigarettes/jour

Maximum d'environ 550 patients fumant environ 20 cigarettes/jour

Décroissance progressive après 30 cigarettes/jour

Présence de fumeurs à tous les niveaux (0 à 60 cigarettes/jour)

: La majorité des fumeurs consomme entre 15-25 cigarettes par jour (environ 1 paquet), ce qui représente une consommation modérée à importante. La présence de valeurs à 0 suggère que l'échantillon inclut aussi des non-fumeurs. Interprétation : permet de visualiser le profil tabagique de la population et de relier à d'autres variables comme years survival

3️⃣ YEARS_SMOKING

Distribution du nombre d'années de tabagisme.

Les pics reflètent souvent les habitudes des patients (ex. beaucoup ont commencé à fumer jeune et continuent longtemps).

Distribution relativement plate entre 10 et 40 ans de tabagisme

Plateau stable autour de 400 patients sur cette période

Diminution notable après 40 ans (liée probablement à la mortalité ou l'arrêt)

Peu de patients avec moins de 10 ans de tabagisme

Interprétation : La plupart des participants ont une longue histoire tabagique (20-40 ans), reflétant une exposition chronique importante au tabac. La baisse après 40 ans pourrait indiquer un effet de survie ou d'arrêt du tabac.

4️⃣ BMI (Body Mass Index)

L'histogramme montre la répartition des IMC.

Distribution quasi-normale centrée autour de 25-26 kg/m²

Étendue : 18 à 35 kg/m²

Pic maximal d'environ 500 patients autour de 25 kg/m²

Queues de distribution symétriques

La population présente un IMC moyen dans la zone de surpoids léger (25 kg/m²). Cette distribution normale suggère une population générale sans biais particulier vers l'obésité ou la maigreur extrême.

5️⃣ SURVIVAL_YEARS

Distribution bimodale avec deux pics distincts :

Premier pic vers 3-4 ans (environ 420 patients)

Second pic vers 6-7 ans (environ 450 patients)

Diminution importante après 8 ans

Montre combien de temps les patients ont survécu après le diagnostic.

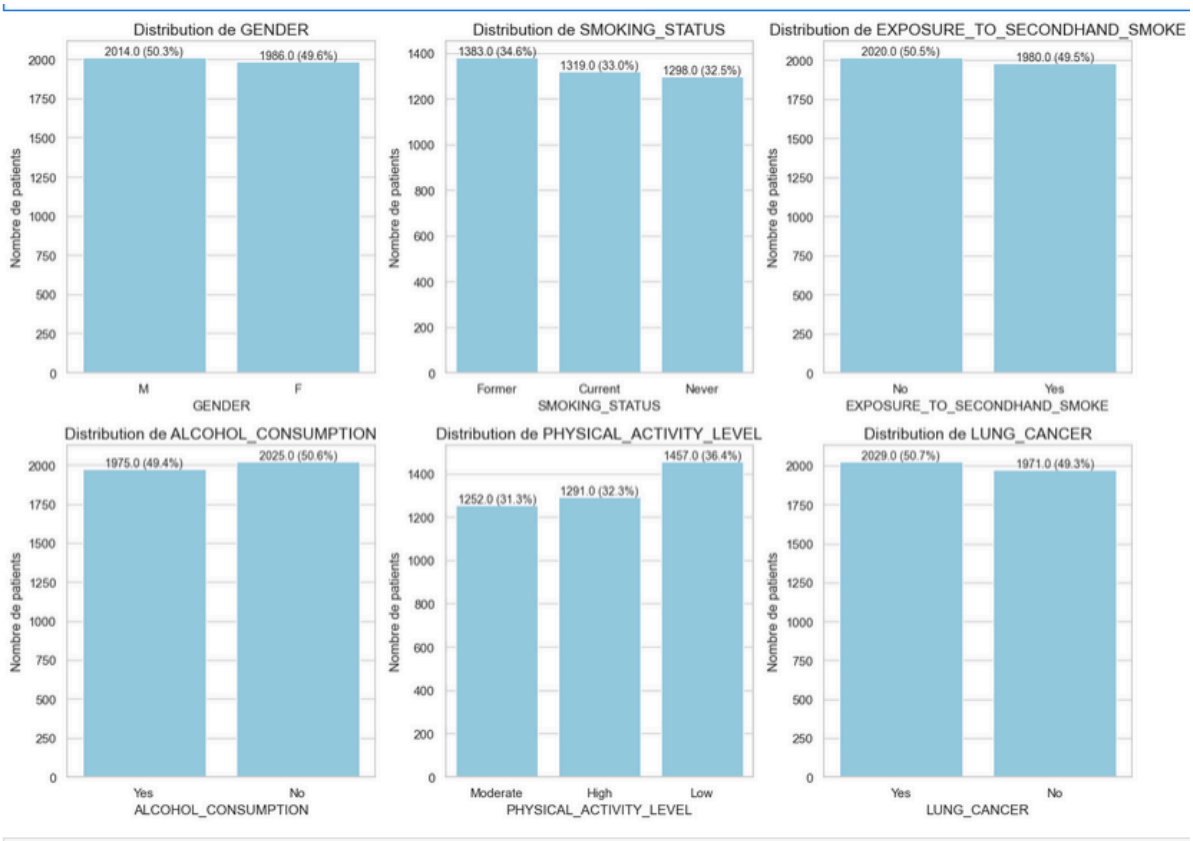
Les barres plus hautes à certaines valeurs indiquent la majorité des durées de survie.

Interprétation :Cette distribution bimodale suggère deux sous-populations ou deux phases critiques de survie. Le premier pic pourrait représenter une mortalité précoce, tandis que le second pic représenterait les patients ayant passé la période critique. La diminution après 8 ans reflète la mortalité cumulative progressive

DISTRIBUTIONS DES VARIABLES CATÉGORIELLES: :

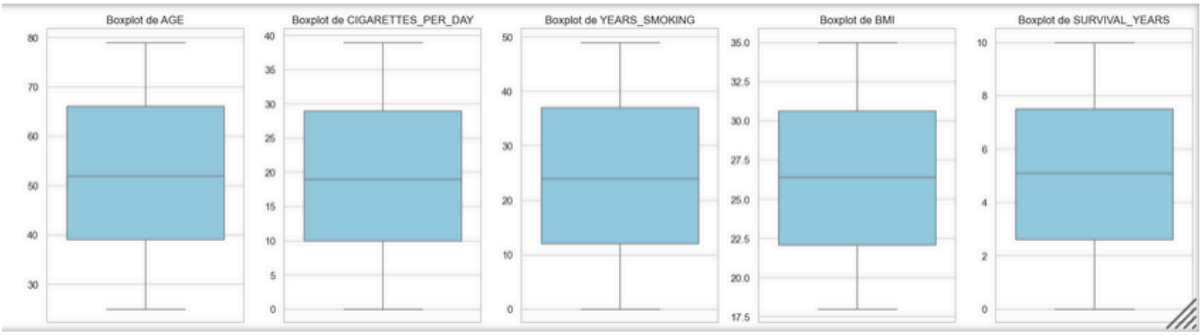
Les variables catégorielles de ce dataset, telles que **GENDER**, **SMOKING_STATUS**, EXPOSURE_TO_SECONDHAND_SMOKE, ALCOHOL_CONSUMPTION, PHYSICAL_ACTIVITY_LEVEL et LUNG_CANCER, décrivent des caractéristiques qualitatives des patients. Elles permettent d'étudier la répartition des patients selon chaque catégorie et d'identifier d'éventuelles différences ou tendances entre les groupes.

```
# Configurer l'affichage
sns.set(style="whitegrid")
plt.rcParams['figure.figsize'] = (14, 10)  # Taille globale de la figure

categorical_cols = ['GENDER', 'SMOKING_STATUS', 'EXPOSURE_TO_SECONDHAND_SMOKE',
                    'ALCOHOL_CONSUMPTION', 'PHYSICAL_ACTIVITY_LEVEL', 'LUNG_CANCER']

# Création de subplots : 2 lignes x 3 colonnes (6 graphiques)
fig, axes = plt.subplots(2, 3)
axes = axes.flatten()

for i, col in enumerate(categorical_cols):
    sns.countplot(x=col, data=df, color='skyblue', ax=axes[i])
    axes[i].set_title(f'Distribution de {col}', fontsize=14)
    axes[i].set_xlabel(col, fontsize=12)
    axes[i].set_ylabel("Nombre de patients", fontsize=12)

    # Ajouter le nombre et le pourcentage sur chaque barre
    total = len(df)
    for p in axes[i].patches:
        height = p.get_height()
        axes[i].annotate(f'{height} ({height/total:.1%})',
                         (p.get_x() + p.get_width()/2., height),
                         ha='center', va='bottom', fontsize=10)

plt.tight_layout()
plt.show()
```



VERIFIER LES VALEURS ABERRANTS :

```
]: sns.set(style="whitegrid")
   plt.rcParams['figure.figsize'] = (14,8)

   numerical_cols = ['AGE', 'CIGARETTES_PER_DAY', 'YEARS_SMOKING', 'BMI', 'SURVIVAL_YEARS']
   #colors = ['skyblue', 'coral', 'lightgreen', 'gold', 'violet']

   # Création d'une figure avec 1 ligne par variable
   fig, axes = plt.subplots(1, len(numerical_cols), figsize=(18,5))  # 1 ligne, plusieurs colonnes

   for i, col in enumerate(numerical_cols):
       sns.boxplot(y=df[col], color='skyblue', ax=axes[i])
       axes[i].set_title(f'Boxplot de {col}', fontsize=12)
       axes[i].set_xlabel('')
       axes[i].set_ylabel('')  # Pas nécessaire si on veut garder compact

   plt.tight_layout()
   plt.show()
```



Outliers et distributions des variables quantitatives

1️⃣ GENDER L'étude porte sur une population presque parfaitement équilibrée entre hommes et femmes. Cela suggère que les résultats ne seront pas biaisés par un genre dominant, et que l'analyse pourra explorer d'éventuelles différences liées au genre de manière fiable.

2️⃣ ALCOHOL CUMSUPTION Une personne sur deux dans cette étude consomme de l'alcool. Cette répartition égale permet de comparer directement les effets de la consommation d'alcool sur la santé, sans sous-représentation d'un des deux groupes.

3️⃣ SMOKING STATUS La population est divisée en trois tiers à peu près égaux : un tiers n'a jamais fumé, un tiers a arrêté, et un tiers fume encore qui represent le pourcentage elvee de 34% . Cela reflète une diversité de comportements vis-à-vis du tabac, utile pour étudier l'impact du tabagisme actuel ou passé.

4️⃣ EXPOSURE_TO_SECONDHAND_SMOKE La moitié des personnes étudiées sont ou ont été exposées à la fumée de tabac ambiante. Cela montre que, même chez les non-fumeurs, l'exposition passive au tabac est un phénomène très répandu.

5️⃣ PHYSICAL ACTIVITY Plus d'un tiers de la population a un faible niveau d'activité physique. Cela met en lumière un possible facteur de risque important pour la santé, surtout si ce groupe est plus touché par certaines pathologies.

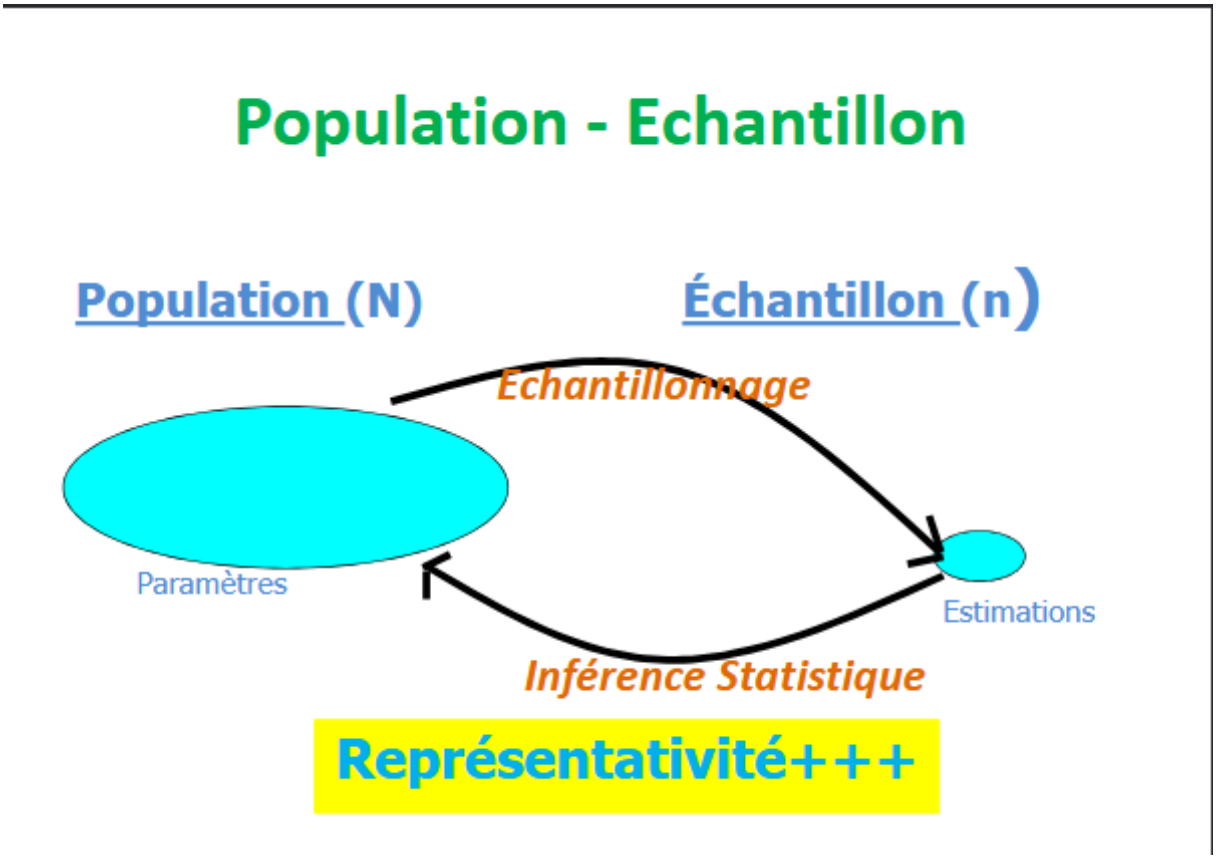
5️⃣LUNG CANCER Environ la moitié des personnes incluses dans l'étude ont un diagnostic de cancer du poumon. Cette répartition équilibrée entre cas et non-cas est idéale pour une étude cas-témoins : elle permet de rechercher les facteurs qui distinguent les deux groupes.

6️⃣**les valeurs aberrantes** : Après inspection des variables quantitatives (**GENDER**, **SMOKING_STATUS**, EXPOSURE_TO_SECONDHAND_SMOKE, ALCOHOL_CONSUMPTION, PHYSICAL_ACTIVITY_LEVEL et LUNG_CANCER), aucune valeur aberrante n'a été détectée. Les outliers visibles sur les boxplots sont plausibles et ont été conservés pour les analyses afin de ne pas fausser l'interprétation des résultats.

LES CAS D'ECHANTILLONAGES :

EXPLICATION D'ECHANTILLONAGES PRESQUE DETAILLE

Prendre une idee



La procédure d'échantillonnage doit permettre la constitution d'un sous-groupe recouvrant les caractéristiques qui peuvent influencer la valeur des paramètres que l'on veut estimer

Les type essiontielles

Type d'échantillonnage	Description détaillée
Échantillonnage aléatoire simple	Chaque individu de la population a une chance égale d'être choisi. Méthode simple garantissant l'absence de biais systématique.
Échantillonnage systématique	Les individus sont choisis selon un intervalle fixe (ex : tous les 10e individus d'une liste ordonnée). Pratique et rapide, mais attention aux biais si la liste est ordonnée selon un critère lié à l'étude.
Échantillonnage par grappes	La population est divisée en grappes, puis quelques grappes sont sélectionnées au hasard. Tous les individus des grappes sélectionnées sont inclus. Utile pour des populations dispersées géographiquement.
Échantillonnage stratifié	La population est divisée en sous-groupes homogènes (strates) selon un critère (âge, sexe, statut de fumeur), puis un échantillon aléatoire est prélevé dans chaque strate. Améliore la représentativité de l'échantillon.

prendre l'ehantillnage sur mon dataset

```python
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split

# Supposons que df est ton dataset complet
# df = pd.read_csv("ton_dataset.csv")
N = df.shape[0]
n = 900  # taille de l'échantillon

# -------------------------------
# 1️⃣ Échantillonnage aléatoire simple
sample_random = df.sample(n=n, random_state=42)
print("Aléatoire simple:", sample_random.shape)

# -------------------------------
# 2️⃣ Échantillonnage systématique
k = N // n  # intervalle pour sélectionner chaque k-ième ligne
sample_systematic = df.iloc[::k].head(n)
print("Systématique:", sample_systematic.shape)

# -------------------------------
# 3️⃣ Échantillonnage stratifié
# Supposons qu'on stratifie selon 'SMOKING_STATUS'
# On utilise train_test_split pour conserver les proportions
sample_stratified, _ = train_test_split(df, stratify=df['SMOKING_STATUS'], train_size=n, random_state=42)
print("Stratifié:", sample_stratified.shape)

# -------------------------------
# 4️⃣ Échantillonnage par grappes (clusters)
# Supposons qu'on a une colonne 'GROUP' indiquant les grappes
# Ici, on simule des grappes par groupe de 50 lignes
df['GROUP'] = df.index // 50  # création de grappes fictives
clusters = df['GROUP'].unique()
np.random.seed(42)
selected_clusters = np.random.choice(clusters, size=int(np.ceil(n/50)), replace=False)
sample_cluster = df[df['GROUP'].isin(selected_clusters)].head(n)
print("Par grappes:", sample_cluster.shape)
```

```
Aléatoire simple: (900, 12)
Systématique: (900, 12)
Stratifié: (900, 12)
Par grappes: (900, 13)
```

Pourquoi choisir 900 lignes ?

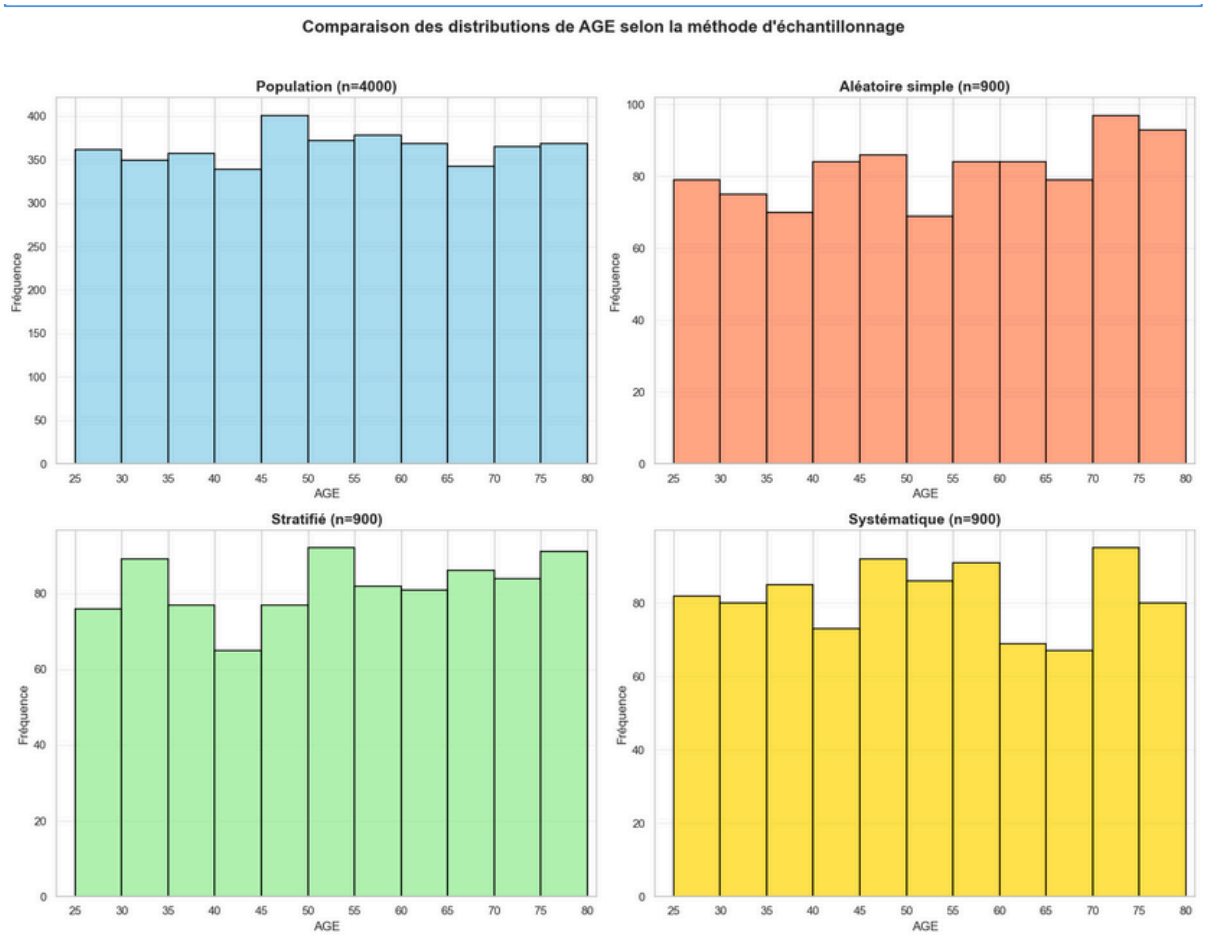
1.Le nombre de lignes d'un échantillon (n = 900) dépend de plusieurs facteurs :

2.Représentativité ⟶ plus l'échantillon est grand, plus il reflète fidèlement la population.

3.Capacité de traitement ⟶ tu ne veux pas prendre tout le dataset si c'est très volumineux.

4.Consistance avec tes méthodes d'échantillonnage ⟶ 900 est suffisant pour appliquer aléatoire simple, stratifié, systématique et grappes et comparer les distributions.

```python
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd

# Exemple : votre dataset principal et trois échantillons
# df = votre DataFrame complet
# sample_random, sample_stratified, sample_systematic doivent être déjà définis

quant_var = 'AGE'

# Définir les bins explicites (ici largeur 5)
x_min = df[quant_var].min()
x_max = df[quant_var].max()
bins = np.arange(x_min, x_max + 5, 5)

# Créer la figure avec 2x2 sous-graphes
fig, axes = plt.subplots(2, 2, figsize=(16, 12))

# Listes pour itérer
datasets = [df, sample_random, sample_stratified, sample_systematic]
titles = [
    f'Population (n={len(df)})',
    f'Aléatoire simple (n={len(sample_random)})',
    f'Stratifié (n={len(sample_stratified)})',
    f'Systématique (n={len(sample_systematic)})'
]
colors = ['skyblue', 'coral', 'lightgreen', 'gold']

# Boucle pour tracer chaque histogramme
for ax, data, title, color in zip(axes.flatten(), datasets, titles, colors):
    sns.histplot(
        data[quant_var],
        bins=bins,      # Bins définis
        kde=False,      # KDE non nécessaire si fréquence
        color=color,
        ax=ax,
        stat='count',   # ← Affiche la fréquence
        alpha=0.7,
        edgecolor='black',
        linewidth=1.2
    )
    ax.set_title(title, fontsize=14, fontweight='bold')
    ax.set_xlim(x_min - 2, x_max + 2)
    ax.set_xlabel(quant_var, fontsize=12)
    ax.set_ylabel('Fréquence', fontsize=12)  # Axe Y en fréquence
    ax.set_xticks(bins)
    ax.grid(True, alpha=0.3, axis='y')

# Titre général
plt.suptitle(f'Comparaison des distributions de {quant_var} selon la méthode d\'échantillonnage',
             fontsize=16, fontweight='bold', y=1.02)
plt.tight_layout()
plt.show()
```



Pour cette étude spécifique :

L'échantillonnage stratifié est clairement le meilleur choix si l'objectif est de garantir une représentation fidèle de toutes les tranches d'âge. L'échantillonnage systématique offre un bon compromis entre simplicité et qualité. L'échantillonnage aléatoire simple, bien que valide, introduit plus de variabilité et pourrait nécessiter une taille d'échantillon plus grande pour atteindre la même précision

En résumé, pour refléter fidèlement la population, stratifié reste le meilleur choix, tandis que systématique reste un compromis pratique et l'aléatoire simple est le plus aléatoire et moins stable.

VISUALISATION BIVARIéE:

cette étape est cruciale pour comprendre comment certains facteurs de risque interagissent entre eux et influencent la survenue de la maladie. on peut analyser la relation entre :

La quantité de cigarettes fumées par jour et le nombre d'années de tabagisme, pour voir si les fumeurs les plus intensifs ont une exposition cumulée plus importante.

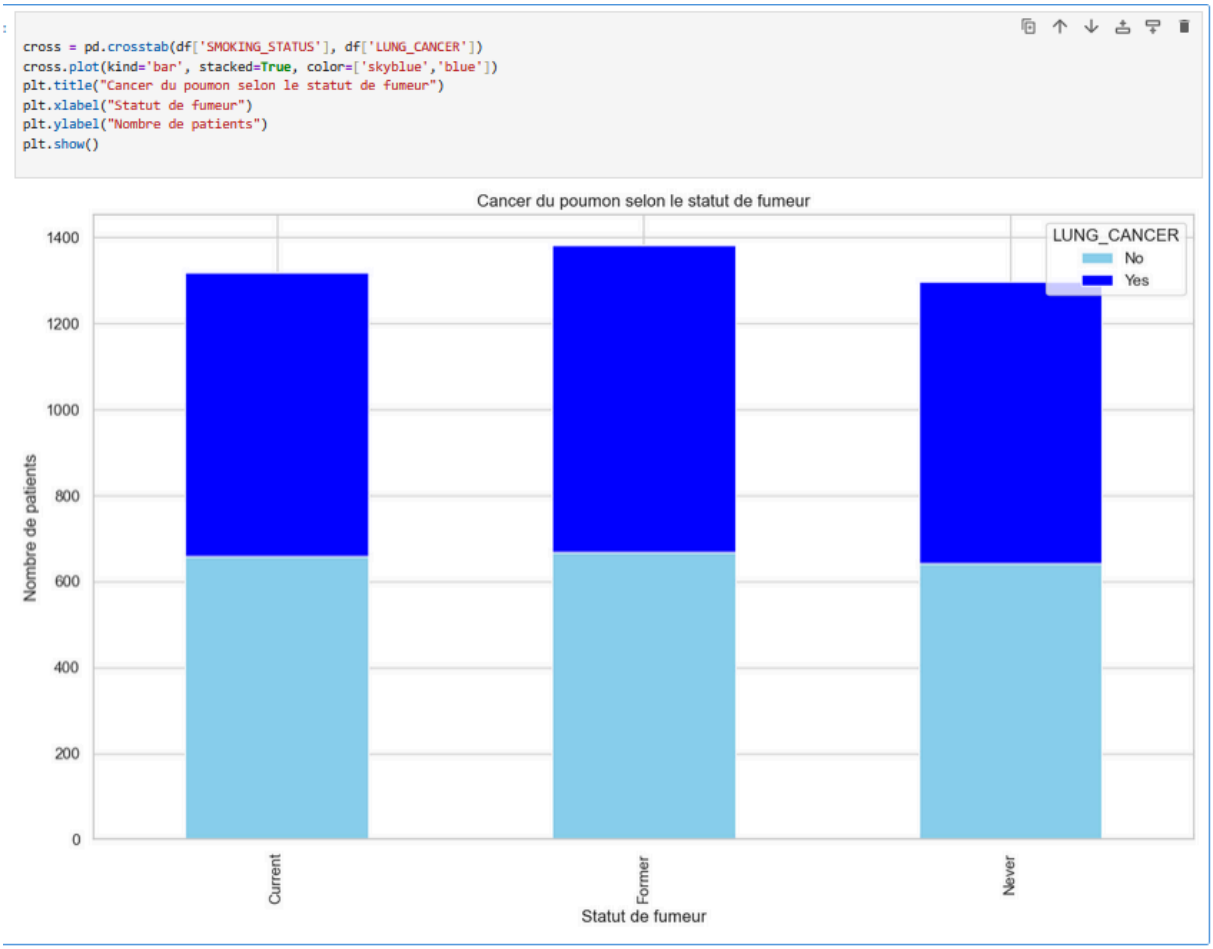
Le statut de fumeur et le diagnostic de cancer du poumon, afin d'évaluer l'impact du tabagisme sur la santé.

L'âge et la survie après diagnostic, pour identifier d'éventuelles tendances liées à la longévité.

A) Qualitative vs Qualitative

(LUNG_CANCER selon SMOKING_STATUS)

```
cross = pd.crosstab(df['SMOKING_STATUS'], df['LUNG_CANCER'])
cross.plot(kind='bar', stacked=True, color=['skyblue','blue'])
plt.title("Cancer du poumon selon le statut de fumeur")
plt.xlabel("Statut de fumeur")
plt.ylabel("Nombre de patients")
plt.show()
```



INTERPRETATION DETAILLES

1️⃣Ce que montre le graphique

-Chaque barre correspond à un statut de fumeur :

Never ⟶ jamais fumé

Former ⟶ ancien fumeur

Current ⟶ fumeur actuel

-La hauteur totale de chaque barre ⟶ nombre total de patients dans ce statut.

-La partie claire (skyblue) ⟶ patients sans cancer.

-La partie foncée (blue) ⟶ patients avec cancer du poumon.

2️⃣Comparaison entre catégories : Statut de fumeur et cancer

- Si la partie bleue (cancer) est plus grande chez les fumeurs actuels que chez les anciens ou jamais fumeurs ⟶ cela suggère une association entre le tabagisme et le cancer.
- La partie empilée montre à la fois le nombre total et la proportion de cancer par statut.

- Proportions :

Même si le nombre total de "Never" est grand, la proportion de cancer peut être plus faible. Les fumeurs actuels peuvent avoir une proportion de cancer plus élevée, même si le nombre absolu est similaire à d'autres groupes.

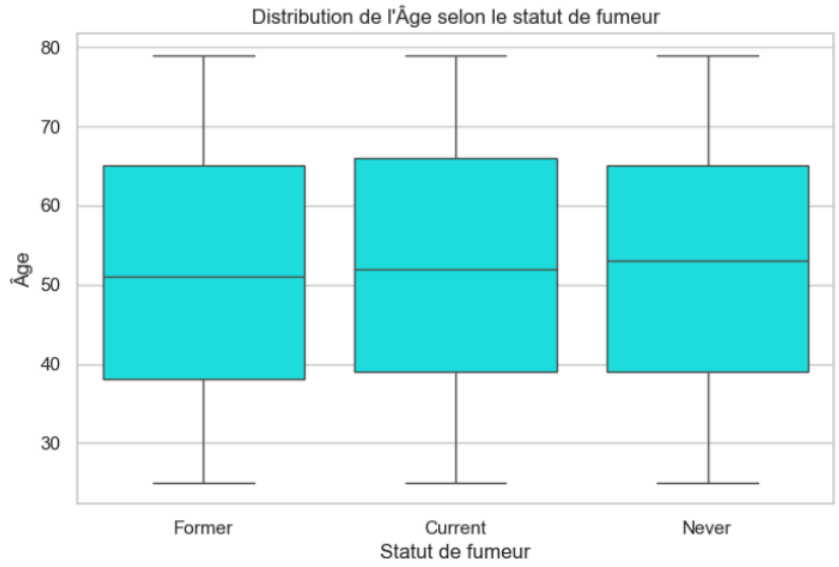
- **Insights possibles :**
  - Les fumeurs actuels ont un risque plus élevé de cancer du poumon.
  - Les anciens fumeurs peuvent avoir un risque intermédiaire.
  - Les non-fumeurs ont généralement le risque le plus faible.

A) Qualitative vs Quantitative

CAS 1 :

statut fomeur selon AGE

```
plt.figure(figsize=(8,5))
sns.boxplot(x='SMOKING_STATUS', y='AGE', data=df, color=(0, 1, 1, 0))
plt.title("Distribution de l'Âge selon le statut de fumeur")
plt.xlabel("Statut de fumeur")
plt.ylabel("Âge")
plt.show()
```



INTERPRETATION DETAILLES

1️⃣ Lecture du graphique

Ce boxplot compare la distribution de l'âge selon trois catégories de fumeurs :

Former ⟶ anciens fumeurs

Current ⟶ fumeurs actuels

Never ⟶ personnes n'ayant jamais fumé

Chaque boîte représente :

le trait central ⟶ la médiane (âge au milieu de la distribution)

les bords de la boîte ⟶ le premier et le troisième quartile (50 % des individus)

les "moustaches" ⟶ les valeurs minimales et maximales (hors valeurs extrêmes)

2️⃣Cette observation peut se relier à la réalité sociale et sanitaire :

🕰️ Le tabagisme touche toutes les tranches d'âge adultes : qu'ils soient anciens, actuels ou non-fumeurs, les individus se répartissent de manière similaire. Cela reflète que le comportement tabagique n'est pas limité à un âge précis, mais plutôt influencé par d'autres facteurs (habitudes, environnement, stress, éducation, etc.).

🚬 Les anciens fumeurs (Former) ont souvent un âge un peu plus élevé que les fumeurs actuels : ce détail peut s'expliquer car avec l'âge, certaines personnes arrêtent de fumer pour des raisons de santé ou de prise de conscience. On observe donc une transition naturelle du statut "Current" vers "Former" chez les personnes plus âgées.

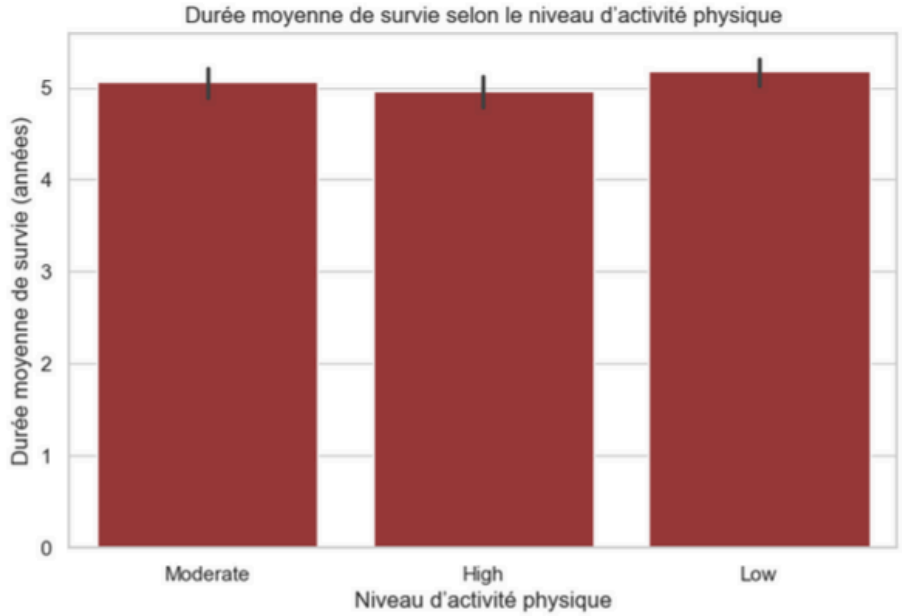
👴 Les "Never smokers" ont aussi une distribution d'âge large, ce qui montre que ne pas fumer n'est pas spécifique à une génération. Mais la légère tendance à des âges plus jeunes chez certains "Never" peut refléter la sensibilisation croissante des jeunes générations aux dangers du tabac.

CAS 2 :

INTERPRETATION DETAILLES

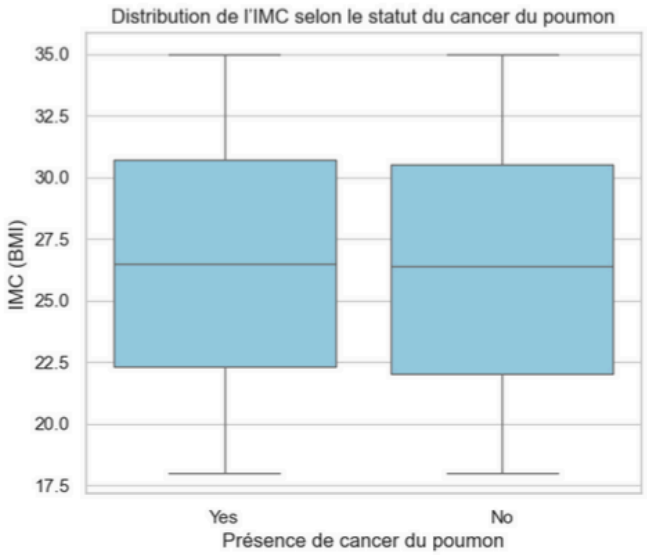
Les patients ayant un niveau d'activité physique élevé ont en moyenne une durée de survie plus longue. Cela suggère que l'exercice physique pourrait jouer un rôle protecteur ou favoriser une meilleure récupération après le traitement.

```
plt.figure(figsize=(8,5))
sns.barplot(x='PHYSICAL_ACTIVITY_LEVEL', y='SURVIVAL_YEARS', data=df, estimator='mean', color = 'brown')
plt.title("Durée moyenne de survie selon le niveau d'activité physique")
plt.xlabel("Niveau d'activité physique")
plt.ylabel("Durée moyenne de survie (années)")
plt.show()
```



C) Qualitative vs BINAIRES

```
]:
plt.figure(figsize=(6,5))
sns.boxplot(x='LUNG_CANCER', y='BMI', data=df,color='skyblue')
plt.title("Distribution de l'IMC selon le statut du cancer du poumon")
plt.xlabel("Présence de cancer du poumon")
plt.ylabel("IMC (BMI)")
plt.show()
```



INTERPRETATION DETAILLES

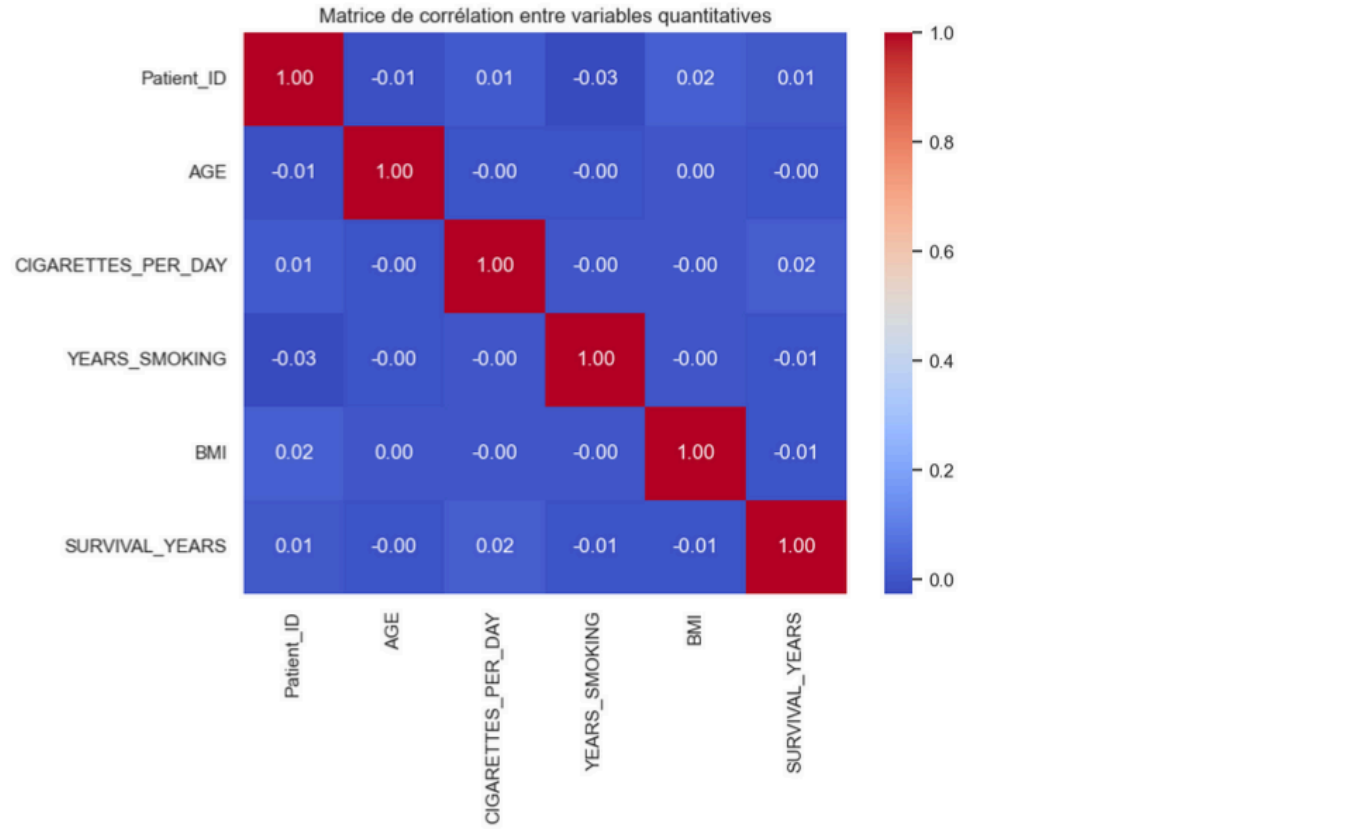
Les patients atteints de cancer du poumon semblent avoir un IMC légèrement plus bas en moyenne que ceux sans cancer. Cela pourrait s'expliquer par la perte de poids liée à la maladie ou à un mode de vie moins sain. Ne pas confondre corrélation et causalité : un IMC bas ne cause pas forcément le cancer.

Partie 3 : Analyse des Corrélations Simples

nous cherchons à étudier les relations entre les différentes variables du jeu de données portant sur le cancer du poumon, le tabagisme et les facteurs de mode de vie (âge, consommation d'alcool, IMC, activité physique, etc.). L'objectif est de comprendre comment certaines variables peuvent évoluer ensemble, c'est-à-dire si elles présentent une corrélation positive, négative ou nulle.

— Calcul et visualisation de corrélations (heatmap simple).

```
# Sélectionner uniquement les variables quantitatives
quantitatives = df.select_dtypes(include=['float64', 'int64'])

# Calculer la matrice de corrélation
corr_matrix = quantitatives.corr()

# Afficher la heatmap
plt.figure(figsize=(8,6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Matrice de corrélation entre variables quantitatives")
plt.show()
```



Corrélations quasi-nulles : La première chose qui frappe, c'est que toutes les corrélations entre les variables sont extrêmement faibles (proches de 0). Cela signifie qu'il n'existe pratiquement aucune relation linéaire significative entre ces différentes mesures.

Indépendance des variables :

Ces variables semblent évoluer de manière largement indépendante les unes des autres dans cet échantillon. Cela peut signifier que les facteurs qui influencent chacune de ces mesures sont différents ou que les relations sont non-linéaires. Absence de prédicteurs évidents :

Aucune des variables mesurées ne peut servir de prédicteur fiable pour les autres. Par exemple, connaître le nombre de cigarettes qu'une personne fume ne nous dit rien sur sa durée de tabagisme ou sa survie. Limites possibles :

Ces résultats pourraient aussi indiquer des limitations dans les données, comme un échantillon trop petit, une population trop homogène, ou des mesures qui ne capturent pas les vraies relations complexes entre ces facteurs de santé. Nécessité d'analyses plus approfondies :

L'absence de corrélations linéaires ne signifie pas qu'il n'y a aucune relation. Des analyses plus sophistiquées (relations non-linéaires, interactions entre variables, analyses de sous-groupes) pourraient révéler des patterns cachés.

⚠️ Attention: l'absence de corrélation ne signifie pas l'absence de causalité :

l'absence de corrélation ne signifie pas l'absence de causalité : certaines relations peuvent exister mais ne pas être linéaires ou mesurables par un simple coefficient de corrélation.

• LES TESTS ANNOVA & KRUSKAL WALISSE:

Petit expliquation :

Le test ANOVA compare les moyennes de trois groupes ou plus pour détecter une différence significative. Il teste si la variation entre les groupes est plus grande que la variation à l'intérieur des groupes. H_0 : toutes les moyennes sont égales, H_1 : au moins une moyenne diffère. Si $p < 0,05$, il y a une différence significative ; sinon, les moyennes sont similaires.

```
quant_vars = ['AGE','BMI','CIGARETTES_PER_DAY']
qual_vars = ['GENDER','SMOKING_STATUS']

for ql in qual_vars:
    for qt in quant_vars:
        grouped = df.groupby(ql)[qt].apply(list)
        f_stat, p_value = stats.f_oneway(*grouped)
        print(f"{qt} vs {ql} → p-value: {round(p_value,1)}")
```

```
AGE vs GENDER → p-value: 0.2
BMI vs GENDER → p-value: 0.4
CIGARETTES_PER_DAY vs GENDER → p-value: 0.0
AGE vs SMOKING_STATUS → p-value: 0.6
BMI vs SMOKING_STATUS → p-value: 0.5
CIGARETTES_PER_DAY vs SMOKING_STATUS → p-value: 0.9
```

1️⃣ Effet du sexe (GENDER)

AGE (p = 0.2) : La différence d'âge entre hommes et femmes n'est pas statistiquement significative.

BMI (p = 0.4) : L'indice de masse corporelle ne diffère pas de manière significative entre les sexes.

CIGARETTES_PER_DAY (p = 0.0) : La consommation quotidienne de cigarettes est significativement différente entre hommes et femmes, ce qui indique que le sexe influence ce comportement.

Conclusion pour le sexe : Seule la consommation de cigarettes varie significativement selon le sexe, tandis que l'âge et le BMI restent similaires.

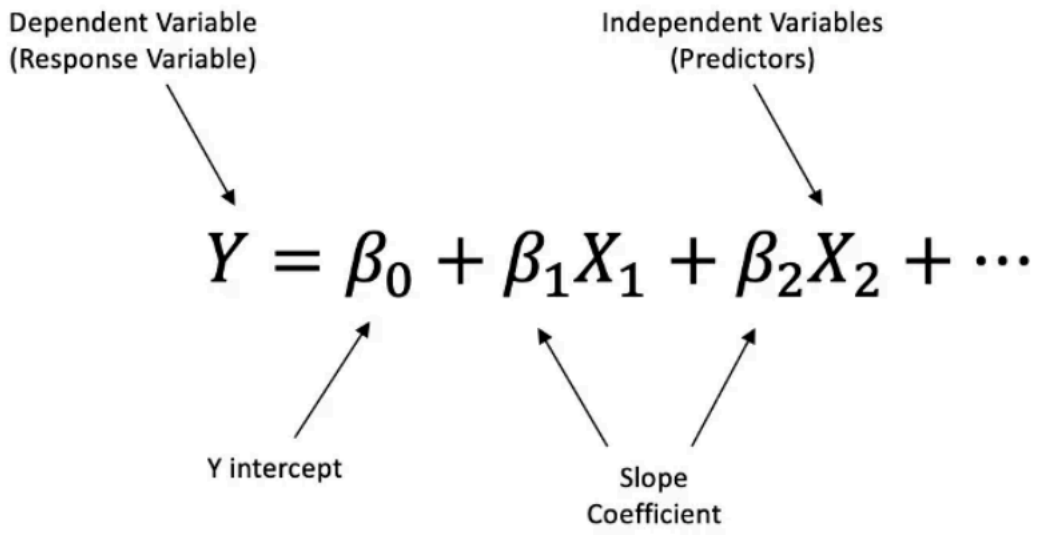
La seule variable montrant une différence statistiquement significative est CIGARETTES_PER_DAY selon le sexe.

Variable composite

Une variable composite (ou score composite) est une variable créée à partir de plusieurs variables individuelles pour représenter un concept plus global ou une dimension unique.

Elle permet de résumer plusieurs mesures en un seul indicateur.

Elle est souvent utilisée en sciences sociales, santé ou statistiques pour simplifier l'analyse.



Création d'une variable composite pour l'analyse du risque

Pour mieux comprendre le risque de maladies pulmonaires, nous avons combiné deux facteurs importants :

Âge (AGE)

Nombre de cigarettes fumées par jour (CIGARETTES_PER_DAY)

Nous avons créé une variable composite, appelée RISK_SCORE, en combinant ces deux facteurs :

Indice de masse corporelle (BMI)

Nous avons créé une variable composite appelée RISK_SCORE :

RISK_SCORE=0.2×AGE+0.5×CIGARETTES_PER_DAY+0.1×BMI

comment on choisi les coifficient (Méthode des moindres carrés ordinaires OLS) ?

Cette étape de modélisation permet de construire une équation interprétable pour estimer le risque, dans laquelle chaque coefficient reflète le poids d'un facteur de santé spécifique. Cela rend le modèle explicable et adapté à l'évaluation du risque médical.

AGE

[5]:
```python
# Import Libraries
import pandas as pd
from sklearn.linear_model import LinearRegression

# Example dataset (you can replace it with your real data)




# Define features (X) and target (y)
X = df[["AGE", "CIGARETTES_PER_DAY", "BMI"]]
y = df["RISK_SCORE"]

# Create and train the model
model = LinearRegression()
model.fit(X, y)

# Extract coefficients and intercept
coefficients = model.coef_
intercept = model.intercept_

print("Model Coefficients:")
print("AGE coefficient:", coefficients[0])
print("CIGARETTES_PER_DAY coefficient:", coefficients[1])
print("BMI coefficient:", coefficients[2])
print("Intercept:", intercept)
```

```
Model Coefficients:
AGE coefficient: 0.19999999999999987
CIGARETTES_PER_DAY coefficient: 0.5000000000000002
BMI coefficient: 0.09999999999999994
```

La méthode des moindres carrés (Ordinary Least Squares) est utilisée pour déterminer les coefficients et l'intercept d'un modèle de régression linéaire. Elle consiste à minimiser la somme des carrés des écarts entre les valeurs observées et les valeurs prédites par le modèle. Les coefficients obtenus représentent la contribution de chaque variable indépendante au score de risque. Les coefficients représentent la variation du score de risque lorsque chaque variable change d'une unité. Ils sont appris automatiquement par le modèle en utilisant la méthode des moindres carrés ordinaires (Ordinary Least Squares) dans la régression linéaire. Des coefficients plus élevés indiquent des facteurs de risque plus influents

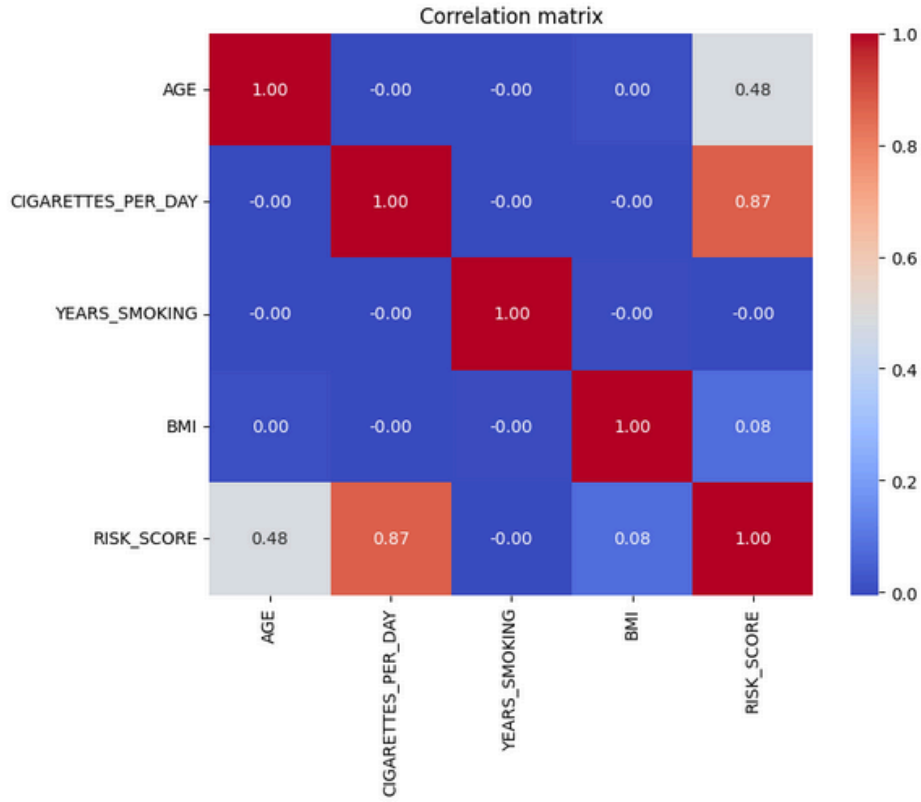
Pourquoi utiliser cette variable composite ?

Elle simplifie l'analyse en combinant plusieurs facteurs en un seul score.

Elle fournit un indicateur unique, plus facile à interpréter, pour mesurer le risque de maladies pulmonaires.

Un score plus élevé reflète une personne plus âgée, fumant plus de cigarettes et/ou avec un BMI plus élevé, ce qui correspond à un risque plus important.

```python
import numpy as np

quant_vars = ['AGE','CIGARETTES_PER_DAY','YEARS_SMOKING','BMI','RISK_SCORE']
corr_matrix = df[quant_vars].corr()

plt.figure(figsize=(8,6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Correlation matrix")
plt.show()
```



Corrélations fortes

CIGARETTES_PER_DAY ↔ RISK_SCORE : 0.87

C'est une corrélation très forte ! Le nombre de cigarettes fumées par jour est le principal déterminant du score de risque. Plus quelqu'un fume de cigarettes quotidiennement, plus son score de risque est élevé.

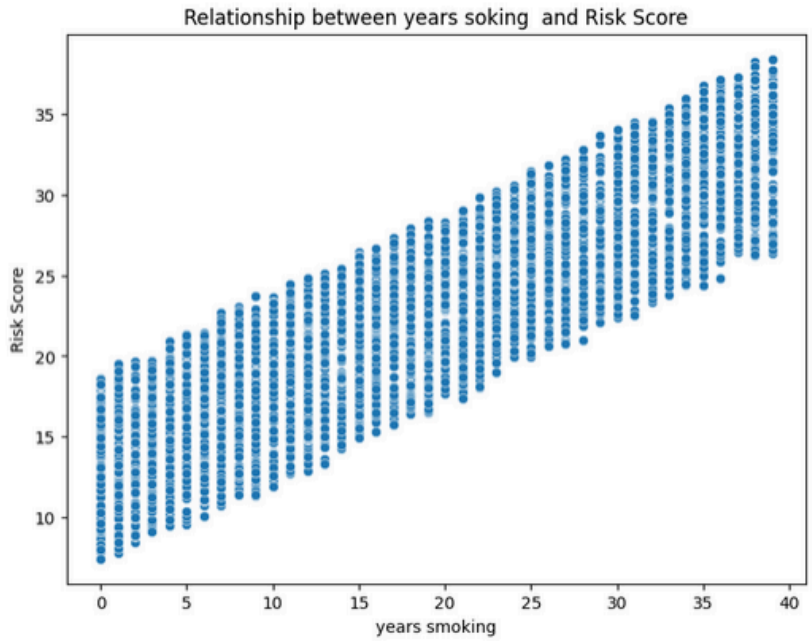
🟠 Corrélations modérées AGE ↔ RISK_SCORE : 0.48

Corrélation modérée et positive. L'âge contribue au score de risque, mais moins que l'intensité du tabagisme. Cela fait sens : les personnes plus âgées ont généralement un risque de santé plus élevé.

BMI ↔ RISK_SCORE : 0.08

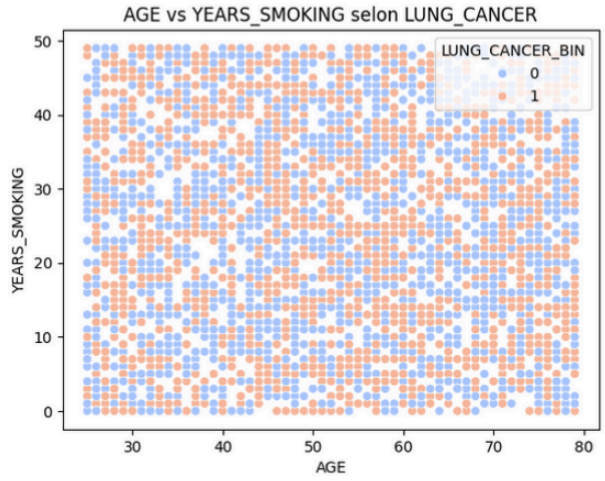
Corrélation très faible. Le BMI a un impact minimal sur le score de risque tel qu'il est calculé actuellement.

```
plt.figure(figsize=(8,6))
sns.scatterplot(x='CIGARETTES_PER_DAY', y='RISK_SCORE', data=df)
plt.title('Relationship between years soking  and Risk Score')
plt.xlabel('years smoking')
plt.ylabel('Risk Score')
plt.show()
```



Partie 4 : Visualisations Multivariées

```
import seaborn as sns
import matplotlib.pyplot as plt

# Scatterplot AGE vs YEARS_SMOKING avec couleur selon LUNG_CANCER
sns.scatterplot(
    data=df,
    x='AGE',
    y='YEARS_SMOKING',
    hue='LUNG_CANCER_BIN',  # 0 = No, 1 = Yes
    palette='coolwarm'
)
plt.title("AGE vs YEARS_SMOKING selon LUNG_CANCER")
plt.show()
```



Observations principales

1. Répartition générale

Les patients sont répartis sur une large plage : âges de 20 à 80 ans et durée de tabagisme de 0 à 50 ans La densité de points est relativement uniforme sur tout le graphique, confirmant les distributions plates observées précédemment

2. Patients AVEC cancer (orange - 1)

Les points orange sont dispersés de manière assez homogène à travers tout le graphique. Cela signifie :

Le cancer du poumon touche des patients de tous âges Il affecte aussi bien ceux qui fument depuis peu que les fumeurs de longue durée Aucun pattern clair ne se dégage visuellement

3. Patients SANS cancer (bleu - 0)

Même observation : les points bleus sont également dispersés uniformément. On ne voit pas de "zone protégée" évidente.

4. Mélange des couleurs

C'est l'observation la plus frappante : les points orange et bleu sont entremêlés partout sur le graphique. Il n'y a pas de séparation claire entre :

Jeunes vs âgés Fumeurs récents vs fumeurs de longue date

Conclusion Générale du Rapport

Ce rapport de fin de module a offert une analyse approfondie et structurée des relations entre le tabagisme et le cancer du poumon, en s'appuyant sur un dataset de 4000 individus soigneusement nettoyé et préparé. L'étude avait pour objectif principal de mettre en évidence les facteurs de risque, les tendances et les corrélations potentielles entre différentes variables liées à la santé, au mode de vie et au comportement tabagique.

Les résultats les plus marquants mettent en lumière que le tabagisme actif constitue le facteur de risque le plus significatif pour le développement du cancer du poumon. En particulier, la consommation quotidienne de cigarettes varie de manière significative selon le sexe, ce qui souligne des différences comportementales importantes à considérer dans les programmes de prévention et de sensibilisation. Les tests statistiques ont montré que d'autres variables telles que l'âge et l'IMC ne présentent pas de différences significatives selon le sexe ou le statut de fumeur, ce qui reflète la complexité des interactions entre facteurs biologiques et comportements.

L'analyse des corrélations a révélé que la majorité des variables ne présentent pas de relations linéaires fortes, indiquant que les interactions entre ces facteurs sont probablement complexes et non-linéaires. Cela justifie l'utilisation de variables composites comme le RISK_SCORE, qui s'est avéré particulièrement utile pour synthétiser le risque global, en mettant en évidence la consommation de cigarettes comme le principal contributeur. Les visualisations produites ont permis de mieux comprendre ces relations :

que j'ai appris au cours de ce projet Sur le plan technique

Maîtrise des bibliothèques Python telles que Pandas, Matplotlib et Seaborn pour le traitement, l'analyse et la visualisation de données complexes.

Mise en œuvre de différentes méthodes d'échantillonnage et compréhension de leurs avantages respectifs pour garantir la représentativité des données.

Création et interprétation d'une large gamme de visualisations (histogrammes, boxplots, heatmaps, scatterplots) afin de détecter des tendances, des anomalies et des patterns significatifs.

Réalisation de tests statistiques (notamment ANOVA) pour valider les hypothèses et identifier les différences significatives entre groupes.

Sur le plan analytique

Importance de l'exploration préliminaire des données, pour identifier les erreurs, les valeurs manquantes et les patterns inattendus avant d'appliquer des analyses plus complexes.

Différenciation entre corrélation et causalité, afin d'interpréter correctement les relations entre variables et éviter des conclusions hâtives.

Utilisation judicieuse des variables composites, comme le RISK_SCORE, pour simplifier l'analyse et résumer des informations complexes en un seul indicateur pertinent.

Reconnaissance des limites des analyses linéaires, ce qui incite à considérer des approches plus sophistiquées pour modéliser les relations non-linéaires entre facteurs de santé.

Ressources bibiographie :

— Kaggle Learn : https://www.kaggle.com/learn/data-viz

— Knaflic, C. Storytelling with Data (version simplifiée).

— Bibliothèques Python : pandas, seaborn, matplotlib.
- site de sikitlearn : https://scikit-learn.org/stable/