Projet n°3 : “Analyse de données de Patients Cancéreux”

Analyse statistique univariée et multivariée

Soutenance de Projet n°3 - Octobre 2021

Ali Naama

Sommaire

I. Description de la problématique et exploration du jeu de données

Description de la problématique

Description du jeu de données :

L'ensemble du jeu de données contient des cas de patients issus d'une étude menée entre 1958 et 1970 à l'hôpital Billings de l'Université de Chicago sur la survie de patientes ayant subi une intervention chirurgicale pour un cancer du sein.

- Nombre d'enregistrements : 306
- Nombre de variables : 4

Objectif du projet :

- Mettre en place une analyse exploratoire des données (EDA : Exploratory Data Analysis) ainsi qu'une analyse statistique univariée et multivariée en langage Python afin de communiquer aux équipes médicales des informations pertinentes
- Prédire en particulier si le patient survivra après 5 ans ou non en fonction de l'âge du patient, de l'année de traitement et du nombre de ganglions lymphatiques positifs

Présentation du jeu de données

Informations sur les variables :

- Âge du patient au moment de l'opération (format numérique)
- Année d'opération du patient (format année — 1900, numérique)
- Nombre de ganglions axillaires positifs détectés (format numérique)
- Statut de survie (attribut de classe)
- 1 = le patient a survécu 5 ans ou plus après intervention chirurgicale
- 2 = le patient est décédé dans les 5 ans après intervention chirurgicale
- Valeurs d'attribut manquantes : aucune

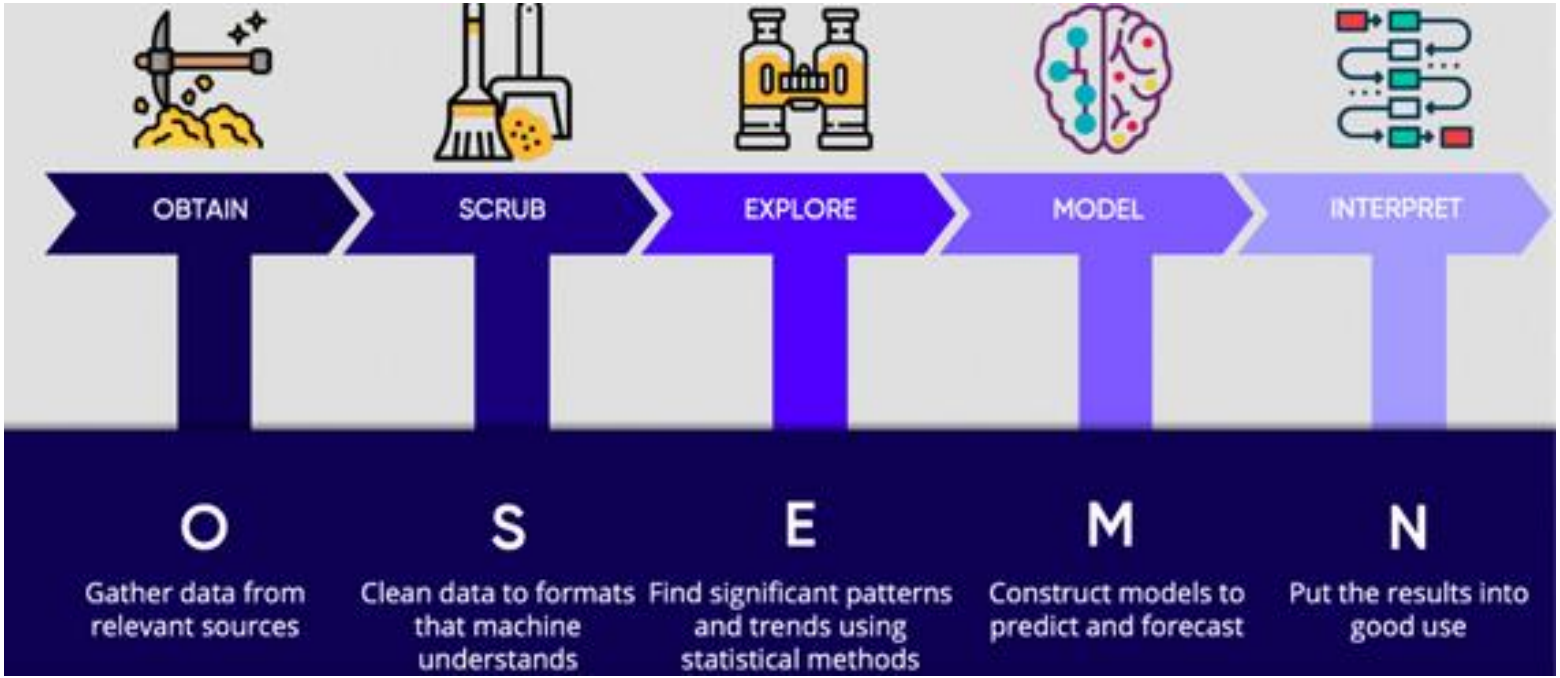
Lien vers la Source des données : Kaggle- habermans-survival-data-set

Outils utilisés pour l'analyse

Nom	Utilisation	Fonctions spécifiques
Anaconda 1.7.2	Gestion de package Gestion d'environnement virtuel	Conda : installation de package via le terminal
Jupyter Notebook 6.1.4	Structurer la démarche Executer code pas à pas Expliquer la démarche (markdown)	
PyCharm 2021.1	Test et développement	IDE Community Edition
Python 3.9.5	Moteur Python Gestionnaire de librairies	Moteur d'exécution
Pandas 1.2.4	Librairie de manipulation de données Représentation des données	Manipulation de Dataframe : création, copie, filtres, tris, description, concaténation, pivotage, autre
Matplotlib 3.4.2 **Seaborn 0.11.2** **Numpy 1.20.3**	Génération de graphiques Gestion des densités de probabilité	Barplot, Scatterplot, lineplot, distplot, heatmap Calcul statistique

II. Analyse Exploratoire

Processus de Data Science



Observations

On notera « habeman », le nom de notre dataframe issu d'une lecture csv via la librairie panda

#	Détail	Syntaxe python
#1	Le fichier ne contient pas d'entête de colonne. Il faut donc les ajouter. Nous ajoutyerons les colonnes suivantes : 'Age', 'Year', 'Positive_Axillary_Nodes', 'Survival_Status'	haberman.head()
#2	Il n'y a pas de valeurs manquantes dans cet ensemble de données.	print(haberman.isnull().sum())
#3	Le type de données de la colonne 'Survival status' est un entier. Il doit être converti en type de données catégoriel. Nous utiliserons les catégories suivantes : « Oui » et « Non »	# Transformation des variables numériques en variables catégorielles haberman['Survival_Status'] = haberman['Survival_Status'].map({1: "Yes", 2: "No"}) haberman['Survival_Status'] = haberman['Survival_Status'].astype('category') print(haberman["Survival_Status"].value_counts())
#4	Synthèse du jeu de données	haberman.describe()

	30	64	1	1.1
0	30	62	3	1
1	30	65	0	1
2	31	59	2	1
3	31	65	4	1
4	33	58	10	1

```
# check for missing values
Age                     0
Year                    0
Positive_Axillary_Nodes 0
Survival_Status         0
```

```
Yes    224
No      81
```

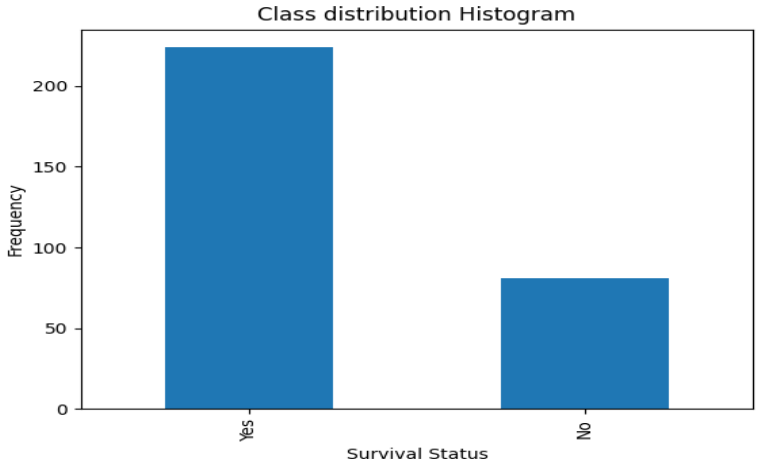
Observations

#	Détail	Syntaxe python

#5	L'âge des patients varie de 30 à 83 ans avec une médiane de 52 ans.
	
	Bien que le nombre maximum de ganglions lymphatiques positifs observés soit de 52, près de 75 % des patients ont moins de 5 ganglions lymphatiques positifs et près de 25 % des patients n'ont pas de ganglions lymphatiques positifs.
	
	L'ensemble de données ne contient qu'un petit nombre d'enregistrements (305).
	
	La colonne cible est déséquilibrée avec 73% des valeurs sont « oui »
		haberman.describe()
		
		# this gives us the distribution of classes in the data set
		print('% of distribution for Survival_Status')
		print(haberman["Survival_Status"].value_counts(1).mul(100).round(1).astype(str) + '%')

```
% of distribution for Survival_Status
Yes    73.4%
No     26.6%
```

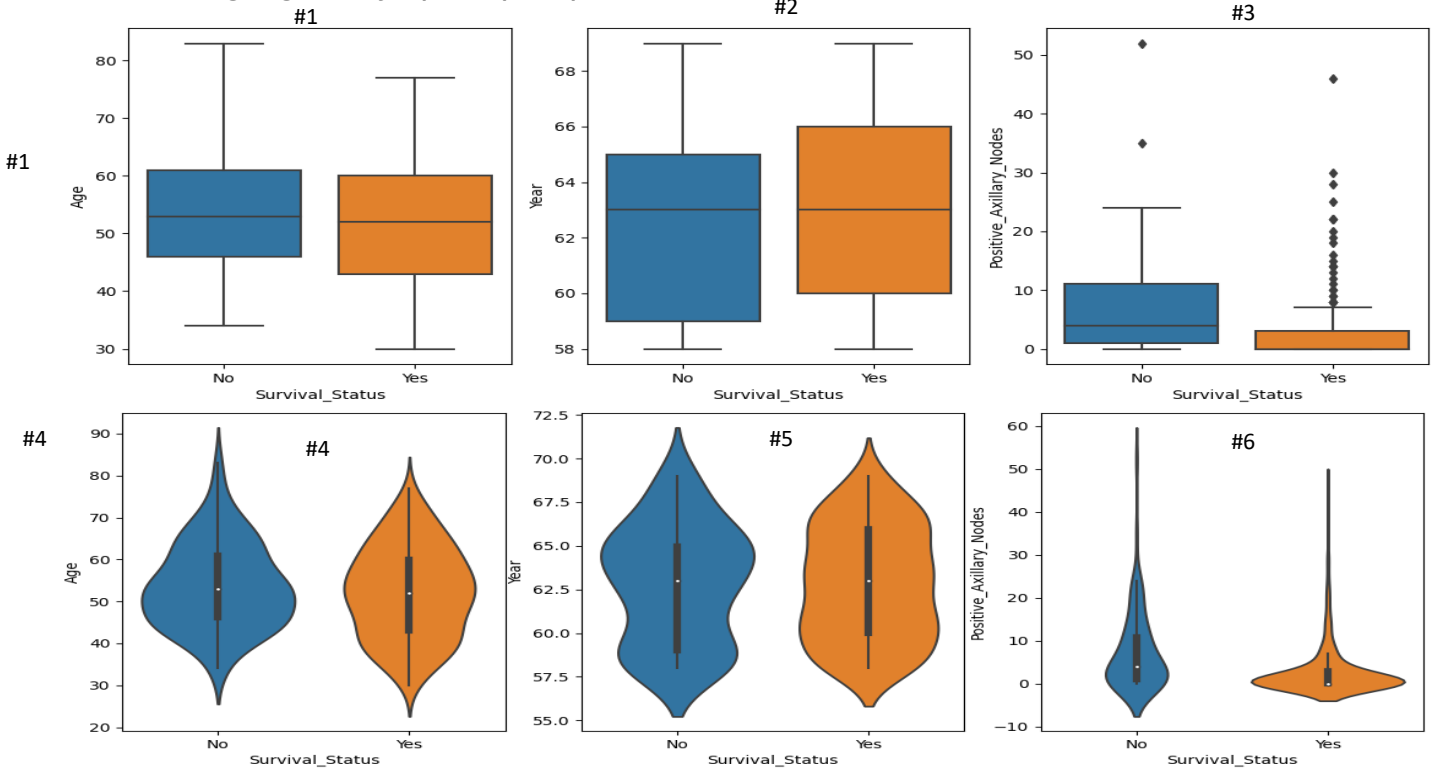
	Age	Year	Positive_Axillary_Nodes
count	305.000000	305.000000	305.000000
mean	52.531148	62.849180	4.036066
std	10.744024	3.254078	7.199370
min	30.000000	58.000000	0.000000
25%	44.000000	60.000000	0.000000
50%	52.000000	63.000000	1.000000
75%	61.000000	66.000000	4.000000
max	83.000000	69.000000	52.000000



III. Analyse statistique univariée

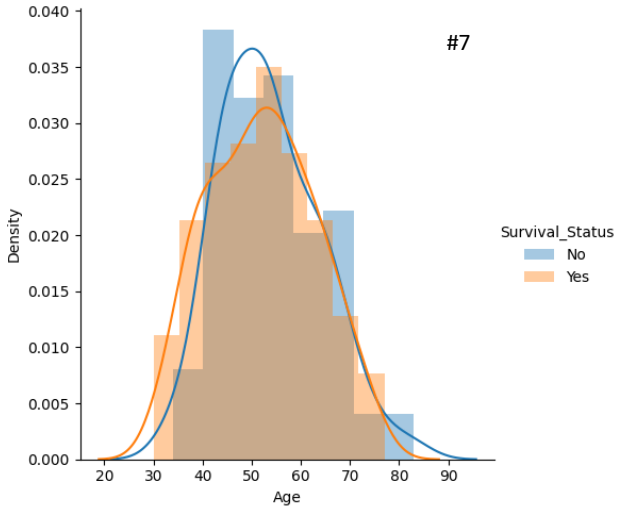
Analyse statistique univariée

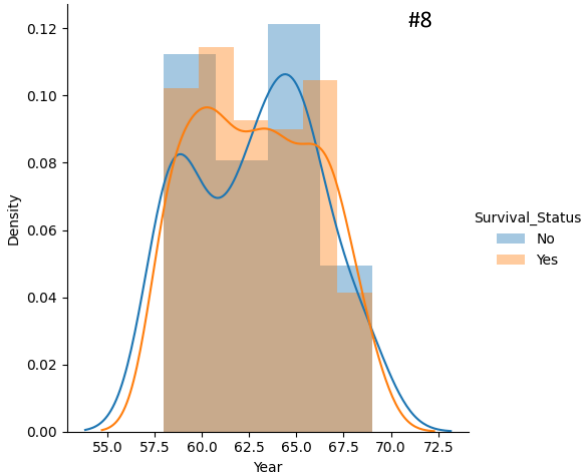
- Étude statistique des modalités d'une seule variable, ou de plusieurs variables considérées indépendamment (Age, Année, Nombre de ganglions), dans le but de décrire l'échantillon de patients cancéreux.
- Prédire si le patient survivra après 5 ans ou non en fonction de l'âge du patient, de l'année de traitement et du nombre de ganglions lymphatiques positifs

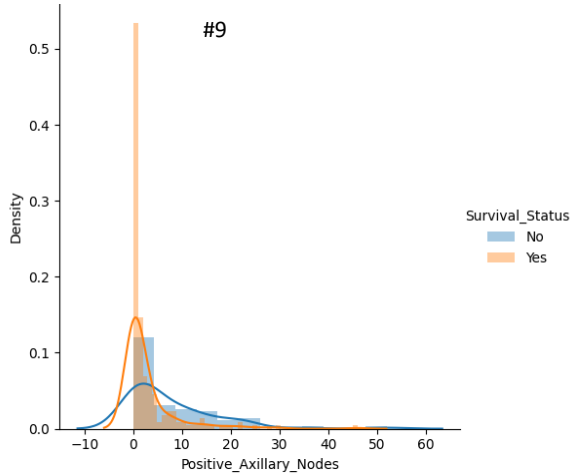


Analyse statistique univariée

- Les graphiques de distribution sont utilisées pour évaluer visuellement la façon dont les points de
- données sont distribués par rapport à la fréquence.
- Habituellement, les points de données sont regroupés via des groupes et la hauteur des barres représentant chaque groupe augmente avec l'augmentation du nombre de points de données
- La fonction de densité de probabilité (**PDF**) est la probabilité que la variable prenne une valeur x. (version lissée de l'histogramme)
- Kernel Density Estimate (**KDE**) est le moyen d'estimer le PDF. L'aire sous la courbe KDE est 1.
- Ici, la hauteur de la barre indique le pourcentage de points de données sous le groupe correspondant







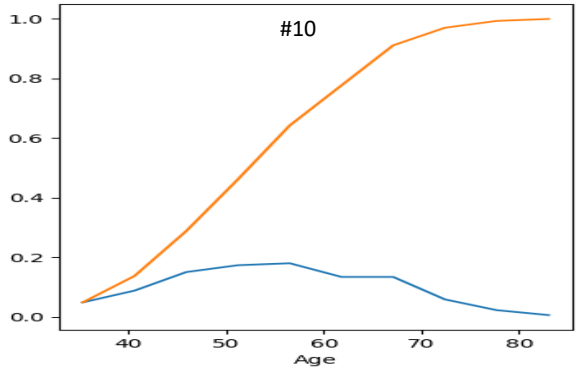
Analyse statistique univariée

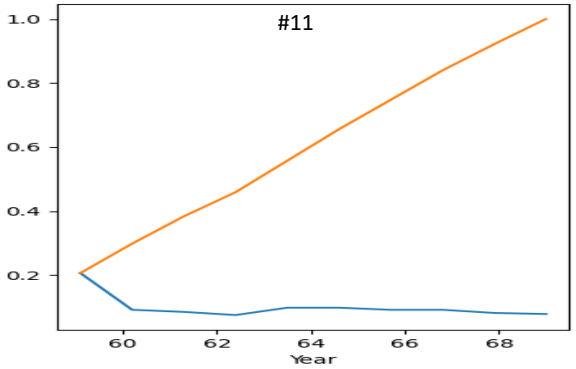
- La fonction de densité de probabilité (**PDF**) est la probabilité que la variable prenne une valeur x. (version lissée de l'histogramme)
- La fonction de distribution cumulative (**CDF**) est la probabilité que la variable prenne une valeur inférieure ou égale à x
- Nous avons positionné un nombre de tranche égal à 10.

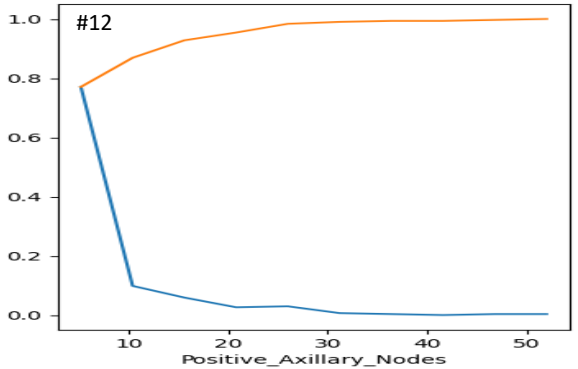
Age	PDF	CDF
30	0,04918033	0,04918033
35,3	0,08852459	0,13770492
40,6	0,15081967	0,28852459
45,9	0,17377049	0,46229508
51,2	0,18032787	0,64262295
56,5	0,13442623	0,77704918
61,8	0,13442623	0,91147541
67,1	0,05901639	0,9704918
72,4	0,02295082	0,99344262
77,7	0,00655738	1
83	0	1

Year	DF	CDF
58	0,20655738	0,20655738
59,1	0,09180328	0,29836066
60,2	0,0852459	0,38360656
61,3	0,07540984	0,45901639
62,4	0,09836066	0,55737705
63,5	0,09836066	0,6557377
64,6	0,09180328	0,74754098
65,7	0,09180328	0,83934426
66,8	0,08196721	0,92131148
67,9	0,07868852	1
69	0	1

Positive_Axillary_Nodes	PDF	CDF
0	0,7704918	0,7704918
5,2	0,09836066	0,86885246
10,4	0,05901639	0,92786885
15,6	0,02622951	0,95409836
20,8	0,0295082	0,98360656
26	0,00655738	0,99016393
31,2	0,00327869	0,99344262
36,4	0	0,99344262
41,6	0,00327869	0,99672131
46,8	0,00327869	1
52	0	1

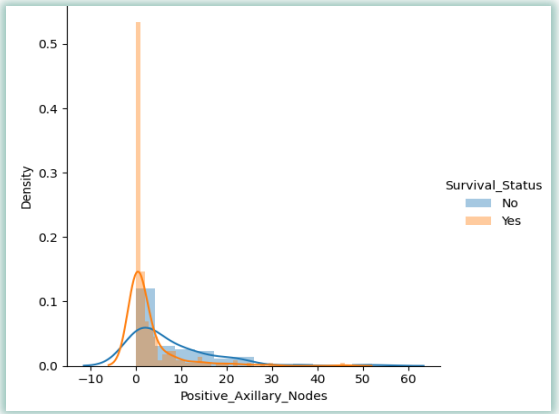


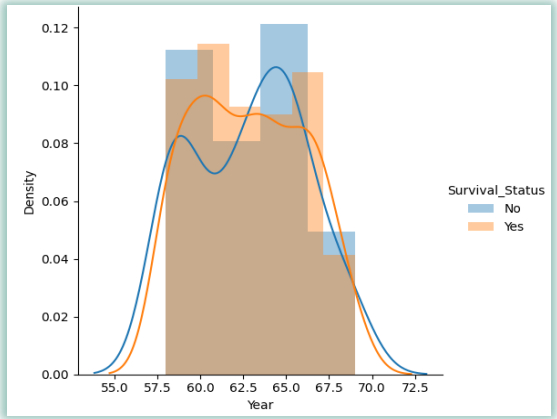




Analyse statistique univariée

- Graphiques # 9 et #12 :
  - Le nombre de ganglions lymphatiques positifs des survivants est très dense autour de 0 à 5.
  - Près de 80 % des patients ont un nombre inférieur ou égal à 5 ganglions lymphatiques positifs.
- Graphiques # 8 et #11 :
  - Les patients traités après 1966 ont un peu plus de chances de survivre que les autres.
  - Les patients traités avant 1959 ont un peu moins de chances de survivre que les autres.

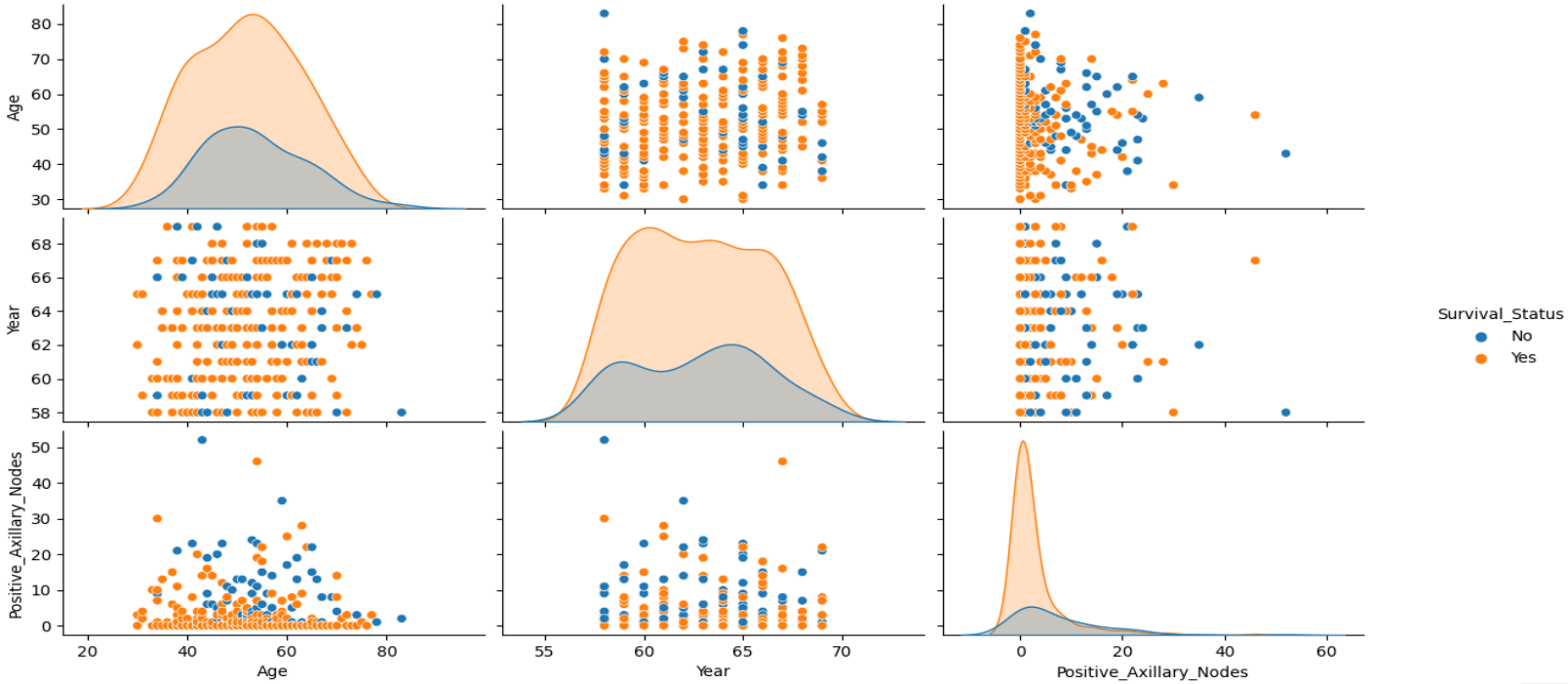




IV. Analyse statistique multivariée

Analyse statistique multivariée

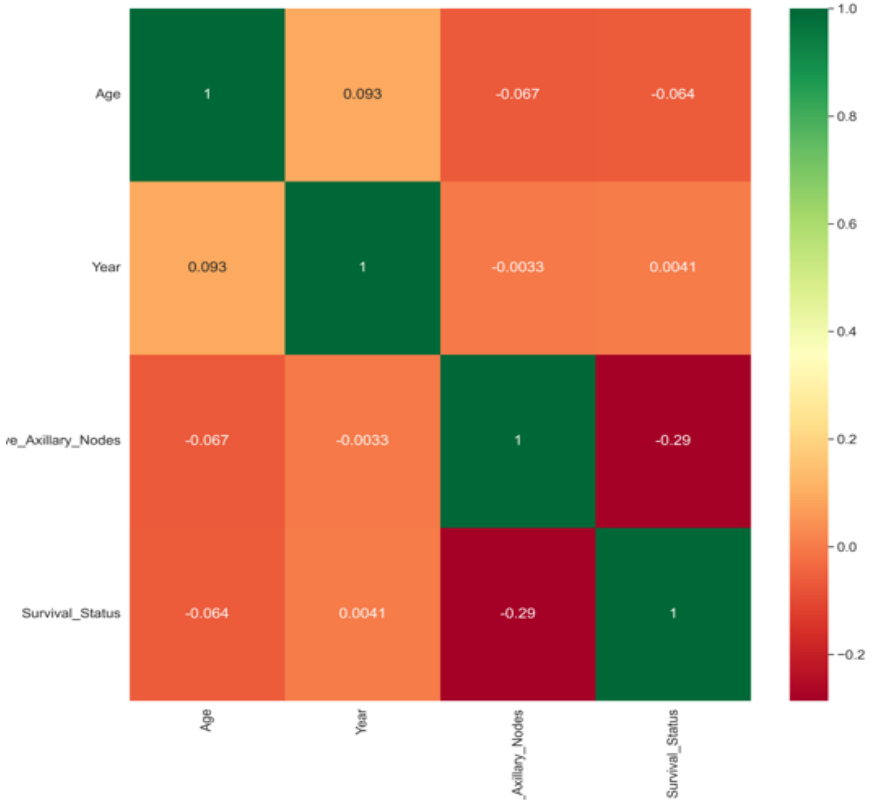
- Visualisons la relation entre deux variables, en particulier entre la variable Nombre de ganglions et les deux autres variables : Age, Année
- En dispersant les points de données entre la variable Year et Nodes, nous pouvons voir la meilleure séparation entre les deux classes que les autres nuages de points.



Analyse statistique multivariée

- Visualisons la matrice de covariance.
- En effet, un bon moyen de vérifier rapidement les corrélations entre les colonnes consiste à visualiser la matrice de corrélation sous forme de carte thermique.
- La valeur du coefficient de corrélation peut prendre n'importe quelle valeur de -1 à 1.

- Nous constatons que la variable du nombre de ganglions agit défavorablement sur la variable statut de survie avec une valeur négative à : -0,29

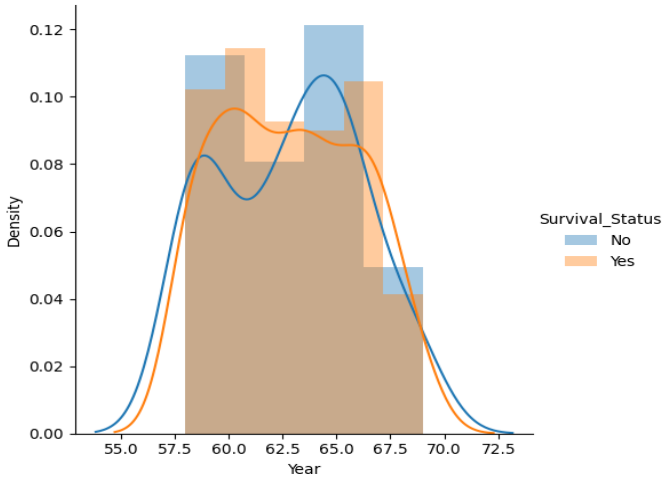



```
# look at the heatmap of the correlation matrix of our dataset
sns.set(font_scale=1.4)
swarm_plot = sns.heatmap(haberman.corr(), annot = True, cmap='RdYlGn')

fig = swarm_plot.get_figure()
fig.set_figwidth(15)
fig.set_figheight(15)
fig.savefig('saving-a-high-resolution-seaborn-plot.png', dpi=300)


# look at the heatmap of the correlation matrix of our dataset
sns.heatmap(haberman.corr(), annot = True)

print(haberman.corr()['Survival_Status']) # numerical correlation matrix
```

V. Résultats et recommandations

Résultats et recommandations

- Grace aux analyses menées, nous pouvons constater que les actions entreprises post 1966 ont permis d'avoir un impact positif sur la baisse de la mortalité. On peut supposer que les actions entreprises sont de type (en se basant sur des actions menées en France sur cette même maladie voir note ci-dessous) :
  - Campagne de préventions auprès de la population
  - Amélioration du dépistage précoce et de la prise en charge
  - Mise en place de nouvelles thérapie



- *: Le cancer du sein est le cancer **le plus fréquent** chez la femme : il représente selon l'Institut national du cancer **33 % des cas de cancers féminins**.*
- ***En 2018, environ 58 500 nouveaux cas de cancer du sein** ont été diagnostiqués en France.*
- *Si cette maladie est encore responsable de plus de **12 100 décès** cette même année, le taux de mortalité qui y est associé diminue régulièrement depuis les années 90.*
- *Cette amélioration s'expliquerait par un meilleur dépistage (60 % des cancers sont aujourd'hui **détectés à un stade précoce**)*
- *mais également par le développement de thérapies toujours plus efficaces. Actuellement, **87 % des patientes sont en vie 5 ans après le diagnostic**.*

VI. Conclusion

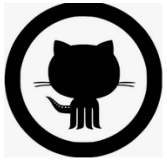
Conclusion

- Grâce à cette étude nous avons pu mettre en pratique les principales librairies nécessaires à l'analyse univariée et multivariée en Python.

- Nous avons pu mettre en pratique une préparation des données et une analyse du jeu de données nécessaire avant analyse statistique plus fine.

- Ce projet, nous permet de nous construire une méthodologie pratique pour la gestion de ce type de cas pour l'avenir et pourra être réutilisé en partie à minima.
- Vous pouvez retrouver le détail du code Python sous Jupyter Notebook via le dépôt Git suivant :

Git



Merci de votre attention