None

Table of Contents

- 1 Introduction
- 2 Régression linéaire
  - 2.1 Decription de la base de données
  - 2.2 Importation des données
  - 2.3 Nettoyage des données
    - 2.3.1 Détection des valeurs manquantes
    - 2.3.2 Détection des doublons
    - 2.3.3 Détection des Outliers
    - 2.3.4 Suppression des Outliers
  - 2.4 Description des données
    - 2.4.1 Distribution des variables
    - 2.4.2 Relation entre la variable cible et les variables numériques
    - 2.4.3 Corrélation entre les variables
  - 2.5 Separation en base de train et test (Data and Target Split)
    - 2.5.1 Entrainement du modéle
    - 2.5.2 Affichage des coefficients
    - 2.5.3 Coefficient de détermination R2
    - 2.5.4 Prédiction du modèle
  - 2.6 Vérification des hypothèses du modèle
    - 2.6.1 Homoscédasticité des résidus
    - 2.6.2 Absence de multi colinéarité entre les variables
    - 2.6.3 Normalité des residus
    - 2.6.4 Indépendance des erreurs
    - 2.6.5 vérification de la Nullité de l'Esperance des erreurs
  - 2.7 Utilisation du modèle pour réaliser des prédictions
- 3 Régression logistique
  - 3.1 Nettoyage des données
    - 3.1.1 Detection des valeurs manquantes
    - 3.1.2 Suppression des valeurs manquantes
    - 3.1.3 Detection des doublons
    - 3.1.4 Detection des Outliers
    - 3.1.5 Suppression des valeurs aberrantes
  - 3.2 Descripition des données
    - 3.2.1 Variables quantitatives
      - 3.2.1.1 Distribution et Boxplot
      - 3.2.1.2 Test Kruskal-wallis afin de confirmer les hypothèses
      - 3.2.1.3 Heatmap de corrélation
    - 3.2.2 Variables categorielles
      - 3.2.2.1 Distribution de la variable cible
      - 3.2.2.2 Distribution des variables catégorielles
      - 3.2.2.3 Test de Chi2 pour confirmer les hypothèses

By Mariam Sylla

- 3.3 Encodage
- 3.4 Sous echantillonnage de la classe majoritaire unsampling
- 3.5 Séparation en base de train et test (Data and Target Split)
- 3.6 Utilisation du modele

Introduction

Dans l'apprentissage supervisé, la régression linéaire permet de prédire une variable cible continue (variable dépendante) grâce à une ou plusieurs variables explicatives (variables indépendantes ou prédictives). En d'autres termes, il s'agit d'établir les relations entre 2 ou plusieurs variables, le but de la régression est d'estimer une valeur (numérique) de sortie à partir des valeurs d'un ensemble de caractéristiques en entrée. Cet algorithme exploite des valeurs numériques pour dégager une tendance ou une évolution prévisible dans le temps. Quant à la régression logistique binaire c'est un modèle statistique permettant d'étudier les relations entre un ensemble de variables qualitatives ou quantitatives Xi et une variable qualitative Y. Il s'agit d'un modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien. Il permet de prédire la probabilité qu'un événement arrive (valeur de 1) ou non (valeur de 0) à partir de l'optimisation des coefficients de régression. Ce résultat varie toujours entre 0 et 1. Dans ce projet nous allons traiter les deux questions (régression linéaire multiple et régression logistique binaire)

Régression linéaire

Decription de la base de données

Le dataset contient des informations sur le prix de vente (Price) des maisons, revenu moyen par région(Income),le nombre de population(Population) de la ville où se trouve la maison, l'age(House_Age) de la maison, le nombre de chambre(Number_Rooms) dans la maison. Nous allons faire une régression linéaire pour prédire le prix de la maison en fonction des différents paramètres. Le dataset a été téléchargé sur Kaggle à travers ce lien. Nous avons supprimé certaines variables et renommer d'autres pour ce projet.

Importation des données

By Mariam Sylla

	Unnamed: 0	Income	House_Age	Number_Rooms	Population	Price
0	0	79545.458574	5.682861	7.009188	23086.800503	1.059034e+06
1	1	79248.642455	6.002900	6.730821	40173.072174	1.505891e+06
2	2	61287.067179	5.865890	8.512727	36882.159400	1.058988e+06
3	3	63345.240046	7.188236	5.586729	34310.242831	1.260617e+06
4	4	59982.197226	5.040555	7.839388	26354.109472	6.309435e+05
...	...	...	...	...	...	...
4995	4995	60567.944140	7.830362	6.137356	22837.361035	1.060194e+06
4996	4996	78491.275435	6.999135	6.576763	25616.115489	1.482618e+06
4997	4997	63390.686886	7.250591	4.805081	33266.145490	1.030730e+06
4998	4998	68001.331235	5.534388	7.130144	42625.620156	1.198657e+06
4999	4999	65510.581804	5.992305	6.792336	46501.283803	1.298950e+06

5000 rows × 6 columns

Nettoyage des données

	Income	House_Age	Number_Rooms	Population	Price
0	79545.458574	5.682861	7.009188	23086.800503	1.059034e+06
1	79248.642455	6.002900	6.730821	40173.072174	1.505891e+06
2	61287.067179	5.865890	8.512727	36882.159400	1.058988e+06
3	63345.240046	7.188236	5.586729	34310.242831	1.260617e+06
4	59982.197226	5.040555	7.839388	26354.109472	6.309435e+05

Détection des valeurs manquantes

	Nombre de valeurs manquantes	Proportion de valeurs manquantes
Income	0	0.0
House_Age	0	0.0
Number_Rooms	0	0.0
Population	0	0.0
Price	0	0.0

Il n'y a pas de de valeurs manquantes

Détection des doublons

```
0
```

By Mariam Sylla

Pas de valeurs doubles

Détection des Outliers

By Mariam Sylla

--------------------------- Income -------------------------

	Income	House_Age	Number_Rooms	Population	Price
12	39033.809237	7.671755	7.250029	39220.361467	1.042814e+06
39	17796.631190	4.949557	6.713905	47162.183643	3.023558e+05
411	36100.444227	5.778489	5.497450	44901.857338	5.995040e+05
558	99629.013581	5.431863	7.351398	36950.739057	1.883481e+06
693	107701.748378	7.143522	8.518608	37619.439929	2.332111e+06
844	39411.652788	4.385845	7.047435	45851.398296	5.394834e+05
962	101928.858060	4.829586	9.039382	22804.991935	1.938866e+06
1096	97548.310413	5.460973	6.609396	39089.415712	2.026303e+06
1271	37971.207566	4.291224	5.807510	33267.767728	3.114052e+04
1459	35963.330809	3.438547	8.264122	24435.777302	1.430274e+05
1597	39294.036523	5.928585	5.960676	43183.516104	7.811375e+05
1734	104702.724257	5.575523	6.932106	22560.527135	1.742432e+06
1891	101144.323930	6.350845	7.231771	35772.524007	2.007556e+06
2025	38139.919045	5.577267	6.348068	45899.738402	7.237501e+05
2092	35608.986237	6.935839	7.827589	20833.007623	4.493316e+05
2242	38868.250311	6.965104	8.966906	25432.076773	7.590447e+05
2300	98468.253641	7.035383	6.629233	50676.312404	2.275455e+06
2597	38734.005216	5.641762	6.297908	38890.892760	4.011486e+05
2719	101599.670580	7.798746	7.480512	37523.864670	2.370231e+06
3069	35454.714659	6.855708	6.018647	59636.402553	1.077806e+06
3144	38571.963670	7.425292	5.723009	47386.793614	9.684116e+05
3183	38122.524488	6.336109	7.762551	38067.552184	8.996093e+05
3483	97881.587279	5.034395	7.575905	37152.799341	1.859161e+06
3541	102881.120902	6.471249	5.693536	21051.531294	1.754938e+06
3798	97669.064491	6.888763	6.739379	43203.271060	2.102244e+06
4087	100741.298585	5.870726	6.644853	26041.487616	1.644923e+06
4400	99317.823145	5.495861	7.182721	50350.352292	2.219724e+06
4449	39777.606906	5.804627	7.147719	38725.424303	6.960145e+05
4716	38530.124478	4.265906	8.026969	67727.229051	1.267987e+06
4744	39653.770031	5.205089	6.951617	40275.599326	3.959013e+05
4844	37908.675863	6.233813	7.252916	39632.079786	8.804028e+05
4855	35797.323122	5.544221	7.795138	24844.200190	2.998630e+05

--------------------------- House_Age -------------------------

	Income	House_Age	Number_Rooms	Population	Price
847	67474.279804	3.278228	7.938421	34971.539554	1.013443e+06
918	68615.767085	8.764786	6.122465	51791.808121	1.836978e+06
1074	65016.223811	2.644304	8.306304	15902.582017	4.145712e+05
1091	75358.482596	8.991399	7.282680	21319.994301	1.740405e+06
1285	71152.756172	3.055370	7.487300	40951.239300	1.002840e+06
1494	57925.044700	3.214868	6.988818	43867.844545	6.338759e+05
1628	71721.421377	2.683043	7.583527	10704.821909	3.954402e+05
1726	81535.149881	3.205828	6.222769	50999.851333	1.166750e+06
1777	61025.390260	8.973441	7.905595	34730.766805	1.468267e+06
1855	68334.782565	8.688434	8.296462	36621.759544	1.817830e+06
1859	72139.646003	3.105751	6.070059	52601.116843	9.458332e+05
2366	78578.343419	3.241716	8.100932	35399.907476	1.248742e+06
2432	56402.542811	3.144894	8.075309	29689.695811	6.964672e+05
2465	61544.583468	2.922736	7.791620	30699.386049	6.777723e+05
2898	57293.427401	9.125283	7.780008	27991.839998	1.561314e+06
3138	77352.637603	9.008900	5.894762	39667.507264	1.898169e+06
3388	70841.996490	8.702960	6.315200	37145.632627	1.621743e+06
3831	64644.898809	3.232059	6.518794	45329.508837	8.633862e+05
3989	65274.139949	9.519088	6.203729	25620.915694	1.399906e+06
4324	77540.894192	3.004717	6.530858	21671.514141	5.557557e+05
4429	69413.613399	3.118802	5.372992	43722.516761	5.781425e+05
4488	85480.437284	2.797215	5.184569	47508.805340	1.033865e+06
4565	57838.744628	2.797619	7.932947	37771.433261	5.899696e+05
4859	73685.403658	8.916093	6.291818	36000.506028	1.798927e+06
4978	80393.339500	8.899713	5.652974	39547.932489	1.910585e+06

By Mariam Sylla

--------------------------- Number_Rooms -------------------------

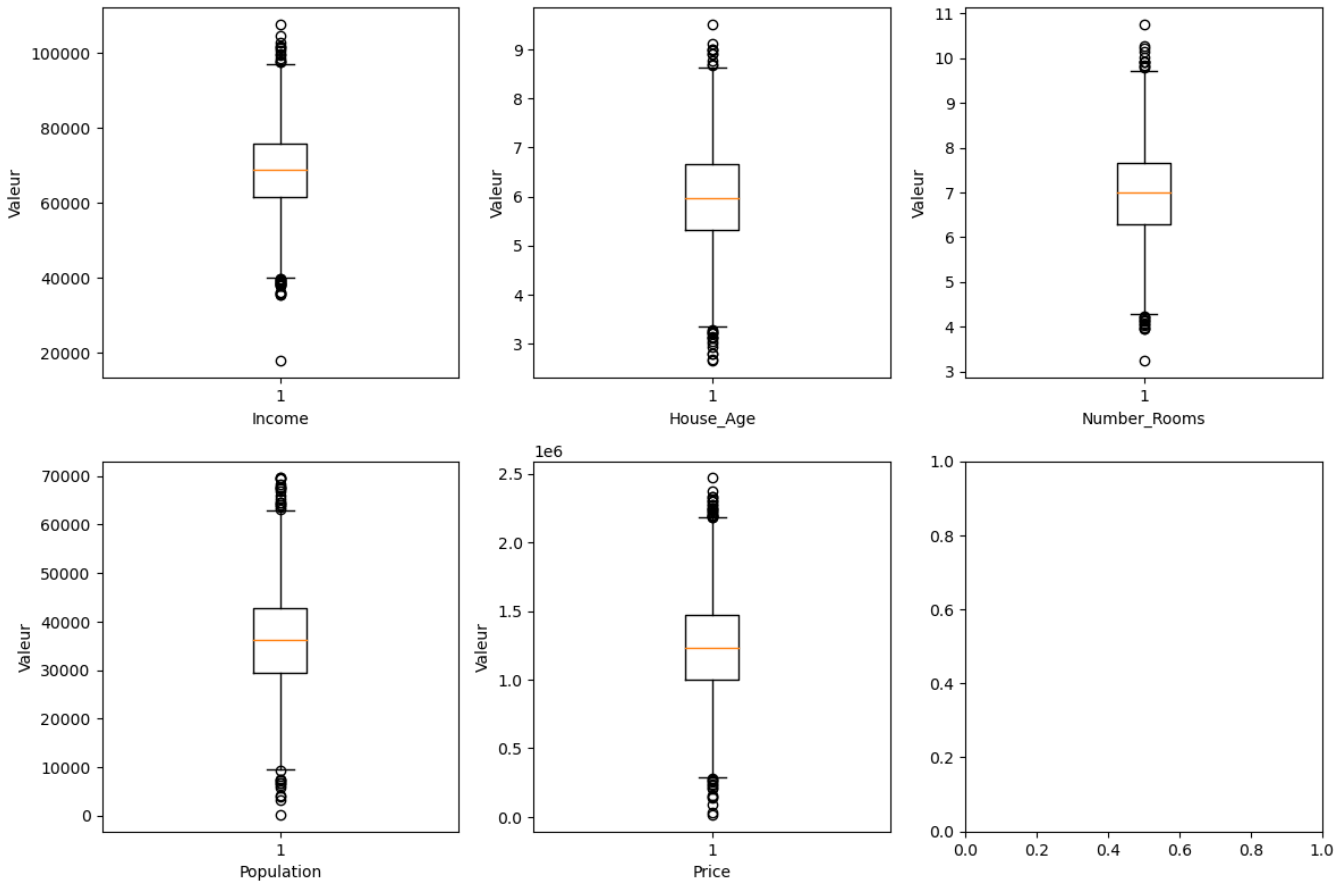
	Income	House_Age	Number_Rooms	Population	Price
28	90499.057451	6.384359	4.242191	33970.164990	1.240764e+06
496	74277.719901	6.987280	3.236194	50233.790310	1.365081e+06
1110	80284.995426	5.029475	4.049321	23457.126397	6.379519e+05
1423	56103.917002	4.725541	4.129733	41330.608556	5.937662e+05
1531	62381.923286	6.777764	9.794898	31391.268468	1.433542e+06
1536	69505.111510	7.352350	10.759588	48112.200174	2.235295e+06
1757	52868.323016	5.823226	3.969632	41849.055084	6.237216e+05
1799	60167.672607	4.590613	3.950973	16811.303292	8.859177e+04
2066	79057.798917	7.634776	10.219902	27142.028199	2.050594e+06
2503	69012.769836	6.981767	9.841095	33069.761581	1.492787e+06
2676	63151.163405	8.381096	9.921520	42266.293372	1.896650e+06
2771	66961.664415	7.412989	3.950225	33423.293292	1.128720e+06
2963	59141.796442	5.631317	4.198677	29386.580449	3.136515e+05
3039	60285.606172	5.142415	9.916528	27370.185718	1.273554e+06
3336	73211.136245	6.501139	10.280022	44198.335811	2.065710e+06
3466	71833.178211	5.002945	9.926147	31750.224524	1.399813e+06
3600	67977.384961	6.978763	4.087718	24246.028058	8.238644e+05
3803	65718.301974	4.907385	4.147431	54798.511451	9.957216e+05
3806	53230.119707	6.274360	10.024375	27360.812223	1.311432e+06
3855	68449.047321	8.390376	10.144988	20145.886508	1.810158e+06
3922	53562.403541	6.323328	4.027931	17964.469901	2.662989e+05
4318	62681.798091	7.341254	4.125278	17454.455217	7.969107e+05
4412	70850.399920	7.306785	4.209620	33845.646950	9.529122e+05
4777	54293.438868	4.772390	9.802010	35079.830306	8.916617e+05

--------------------------- Population -------------------------

	Income	House_Age	Number_Rooms	Population	Price
228	66574.709994	5.550265	6.844150	69575.449464	1.702406e+06
314	68929.158074	7.590878	6.891969	172.610686	8.402729e+05
353	72445.033303	5.488197	6.509449	69553.988327	1.726719e+06
1234	63785.551276	7.196314	6.357874	67353.965204	1.747245e+06
1361	55621.899104	3.735942	6.868291	63184.613147	1.102641e+06
1530	85175.200626	7.750852	7.271163	3285.450538	1.305972e+06
1595	64350.284571	6.761203	7.128710	69592.040236	1.772391e+06
1965	63884.926411	7.722146	5.791958	64566.687380	1.667561e+06
2108	59391.056781	6.492409	8.784536	64543.322446	1.599416e+06
2173	50143.644854	4.230051	7.979250	67601.223558	1.168588e+06
2380	56073.892443	6.576733	6.959056	64149.680213	1.409762e+06
2392	87272.093393	5.025866	7.184765	7522.333138	9.100996e+05
2422	54236.696249	5.643911	8.473199	65857.933322	1.372969e+06
2534	79442.505795	6.300600	5.983205	4114.489353	9.653187e+05
2603	61667.720801	5.593385	7.333572	65184.578469	1.394132e+06
2756	62173.580099	5.098959	5.662268	3883.448164	2.311898e+05
2829	62926.701149	6.500169	6.411871	6248.756080	6.394717e+05
2839	82373.208429	6.327773	7.207986	67701.649795	2.130762e+06
3120	72416.480897	6.162244	7.268954	6805.740783	9.974525e+05
3134	74112.793288	4.412585	8.054622	66995.474049	1.737571e+06
3387	54600.194523	5.608723	5.062462	64180.370801	1.078779e+06
3442	48879.259763	8.147518	7.149646	63620.011963	1.575492e+06
3540	66989.504691	4.842496	6.448919	6821.950228	4.903386e+05
3991	50041.125224	5.981267	8.699555	68311.695822	1.626676e+06
4182	54465.747366	5.754172	6.375873	7234.963521	3.955232e+05
4290	83145.085713	7.190976	9.309902	7360.295191	1.572815e+06
4491	66513.884448	6.778404	8.044497	5727.485885	7.176616e+05
4684	76118.140643	5.285609	7.434269	9193.833182	7.093482e+05
4716	38530.124478	4.265906	8.026969	67727.229051	1.267987e+06
4803	61846.135900	5.057578	7.681141	69621.713378	1.504316e+06

By Mariam Sylla

```
-------------------------- Price -------------------------
            Income  House_Age  Number_Rooms    Population         Price
90     48904.983269   4.844973      5.448956  32960.753070  2.018981e+05
256    91159.418327   6.536045      7.373851  54861.091097  2.298379e+06
263    40366.616291   4.902940      7.617118  16349.365394  1.520719e+05
355    87266.340225   8.248959      7.234261  45161.187677  2.249123e+06
465    90592.469609   7.700132      9.708803  37223.876167  2.469066e+06
622    90890.485814   7.510171      7.595487  45519.256271  2.252243e+06
693   107701.748378   7.143522      8.518608  37619.439929  2.332111e+06
696    49851.134784   4.684996      5.259695  32511.846268  2.832081e+05
715    92280.497474   7.258627      8.222633  38004.145211  2.237778e+06
901    89089.432075   7.146246      9.179994  49782.152070  2.271113e+06
924    83814.101156   8.571797      7.392164  51538.056796  2.330290e+06
990    82915.911433   7.078994      7.882701  50445.647368  2.185480e+06
1208   83936.341967   7.704505      6.193618  55471.783379  2.198565e+06
1248   94733.971275   7.885829      7.162373  46314.690046  2.318286e+06
1271   37971.207566   4.291224      5.807510  33267.767728  3.114052e+04
1356   56654.962390   5.187860      5.336778  25801.965929  2.393199e+05
1459   35963.330809   3.438547      8.264122  24435.777302  1.430274e+05
1485   94670.045916   6.926414      8.340524  39522.909570  2.186195e+06
1516   95997.671100   6.685863      6.993422  47128.880987  2.220799e+06
1536   69505.111510   7.352350     10.759588  48112.200174  2.235295e+06
1578   50926.776634   4.507953      6.154788  33663.669244  2.110180e+05
1661   48735.924512   5.543730      6.091906  19682.347295  1.515271e+05
1799   60167.672607   4.590613      3.950973  16811.303292  8.859177e+04
2300   98468.253641   7.035383      6.629233  50676.312404  2.275455e+06
2538   82859.591647   8.090383      6.927192  60040.547298  2.294648e+06
2719  101599.670580   7.798746      7.480512  37523.864670  2.370231e+06
2756   62173.580099   5.098959      5.662268   3883.448164  2.311898e+05
2795   94292.263625   5.993314      7.278783  53620.890504  2.197437e+06
3091   46367.205859   5.290720      5.181614  26015.296447  2.680508e+05
3212   47320.657205   3.558054      7.006987  15776.618595  1.593866e+04
3502   91046.041784   6.379433      8.811820  37086.621058  2.190339e+06
3922   53562.403541   6.323328      4.027931  17964.469901  2.662989e+05
4129   77701.054052   7.804094      9.287086  49495.064963  2.187326e+06
4400   99317.823145   5.495861      7.182721  50350.352292  2.219724e+06
4451   52588.683645   5.261432      5.053066  32938.355645  2.531857e+05



Variable: Income
Nombre de valeurs aberrantes: 32
Proportion de valeurs aberrantes: 0.64%
------------------------------
Variable: House_Age
Nombre de valeurs aberrantes: 25
Proportion de valeurs aberrantes: 0.50%
------------------------------
Variable: Number_Rooms
Nombre de valeurs aberrantes: 24
Proportion de valeurs aberrantes: 0.48%
------------------------------
Variable: Population
Nombre de valeurs aberrantes: 30
Proportion de valeurs aberrantes: 0.60%
------------------------------
Variable: Price
Nombre de valeurs aberrantes: 35
Proportion de valeurs aberrantes: 0.70%
------------------------------
```



Nous avons des Outliers au niveau des variables "Income", "House_Age", "Number_Rooms","Population" et "Price" que nous allons supprimer.

Suppression des Outliers

	Income	House_Age	Number_Rooms	Population	Price
0	79545.458574	5.682861	7.009188	23086.800503	1.059034e+06
1	79248.642455	6.002900	6.730821	40173.072174	1.505891e+06
2	61287.067179	5.865890	8.512727	36882.159400	1.058988e+06
3	63345.240046	7.188236	5.586729	34310.242831	1.260617e+06
4	59982.197226	5.040555	7.839388	26354.109472	6.309435e+05
...	...	...	...	...	...
4995	60567.944140	7.830362	6.137356	22837.361035	1.060194e+06
4996	78491.275435	6.999135	6.576763	25616.115489	1.482618e+06
4997	63390.686886	7.250591	4.805081	33266.145490	1.030730e+06
4998	68001.331235	5.534388	7.130144	42625.620156	1.198657e+06
4999	65510.581804	5.992305	6.792336	46501.283803	1.298950e+06

4856 rows × 5 columns

By Mariam Sylla

Description des données

skimpy summary

Data Summary

dataframe	Values
Number of rows	4856
Number of columns	5

Data Types

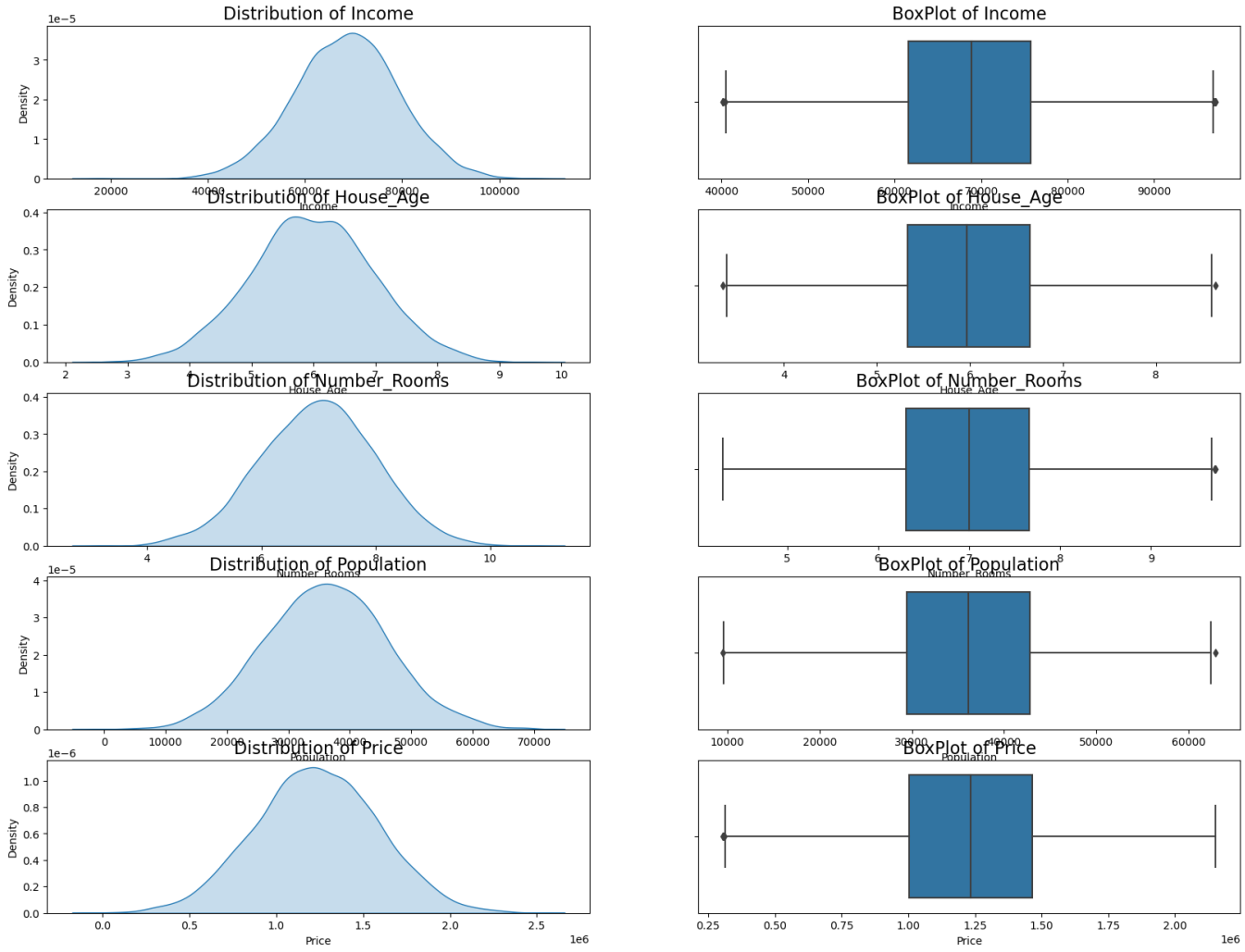
Column Type	Count
float64	5

number

column_name	NA	NA %	mean	sd	p0	p25	p50	p75
Income	0	0	69000	10000	40000	62000	69000	76
House_Age	0	0	6	0.97	3.3	5.3	6	
Number_Rooms	0	0	7	0.98	4.3	6.3	7	
Population	0	0	36000	9600	9500	29000	36000	43
Price	0	0	1200000	340000	300000	1000000	1200000	1500

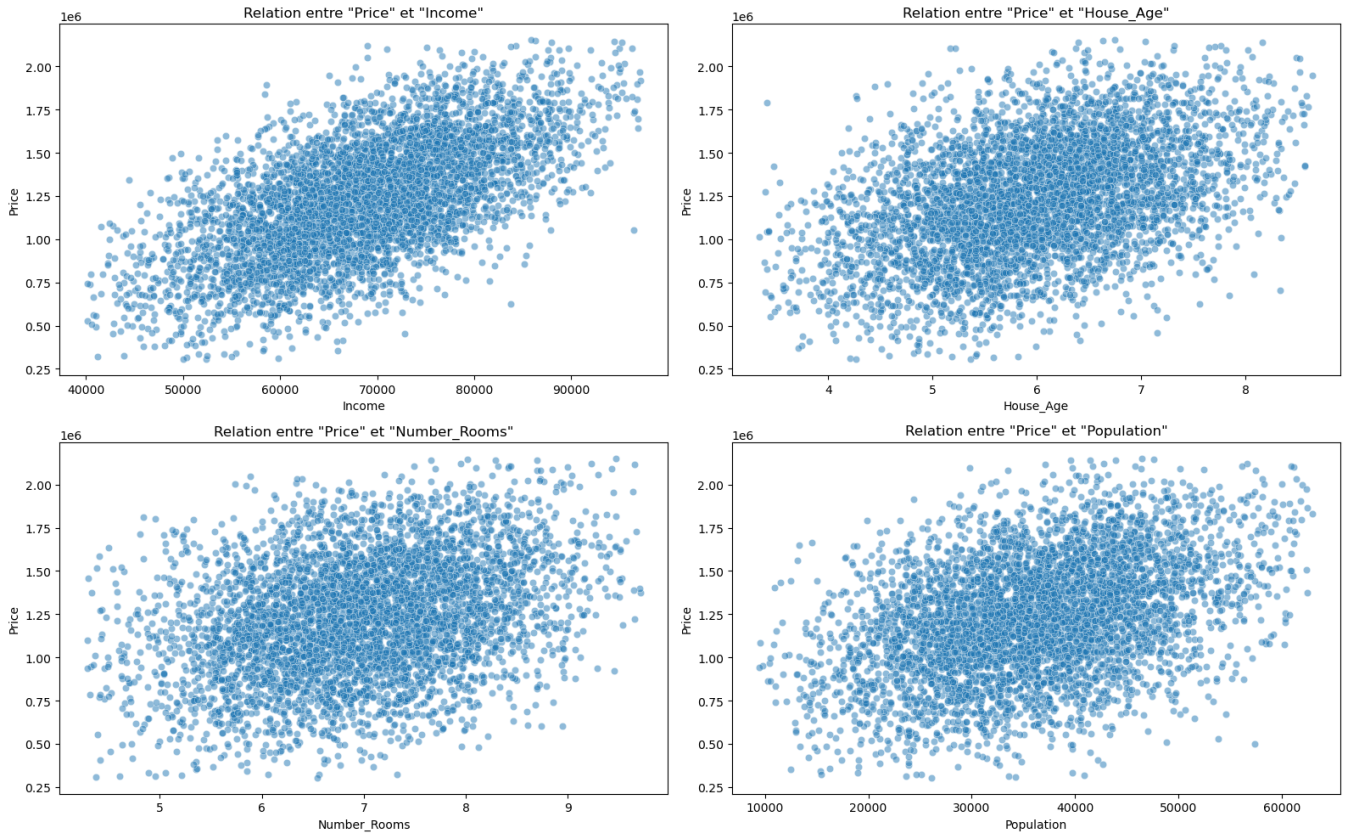
End

Distribution des variables



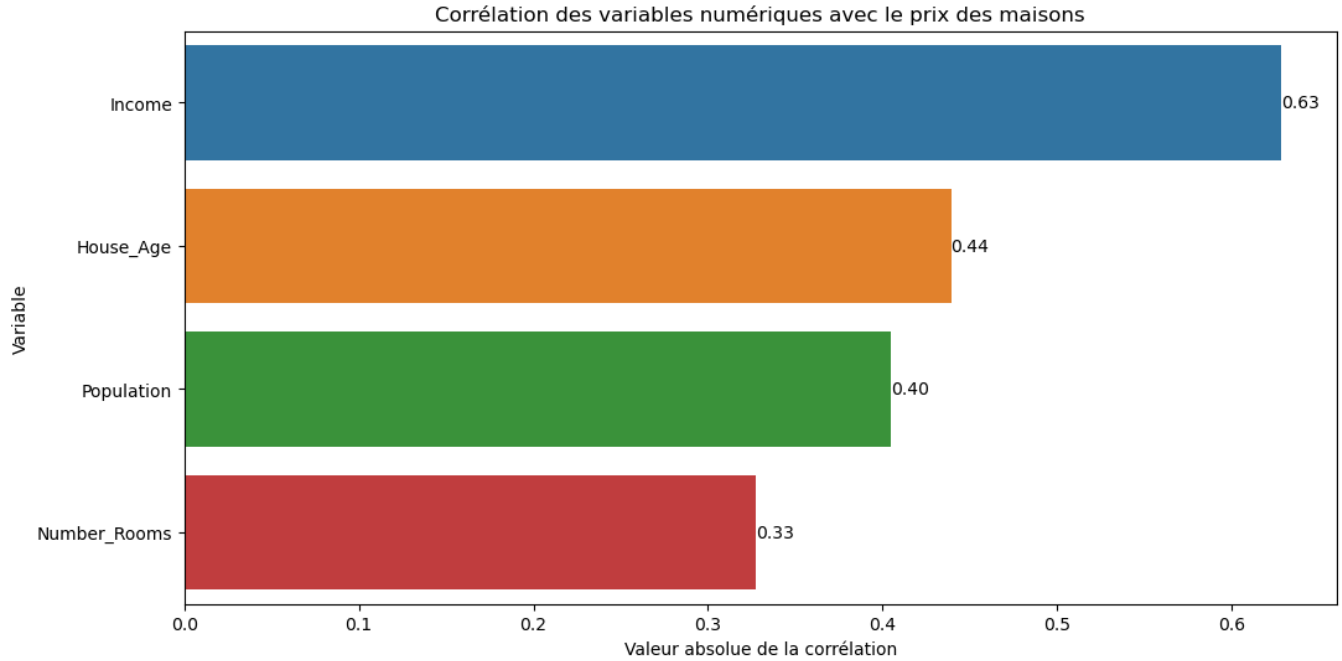
By Mariam Sylla

Relation entre la variable cible et les variables numériques



Toutes les variables ont une relation linéaire avec la variable cible

Corrélation entre les variables



Separation en base de train et test (Data and Target Split)

Entrainement du modéle

By Mariam Sylla

LinearRegression

LinearRegression()

Affichage des coefficients

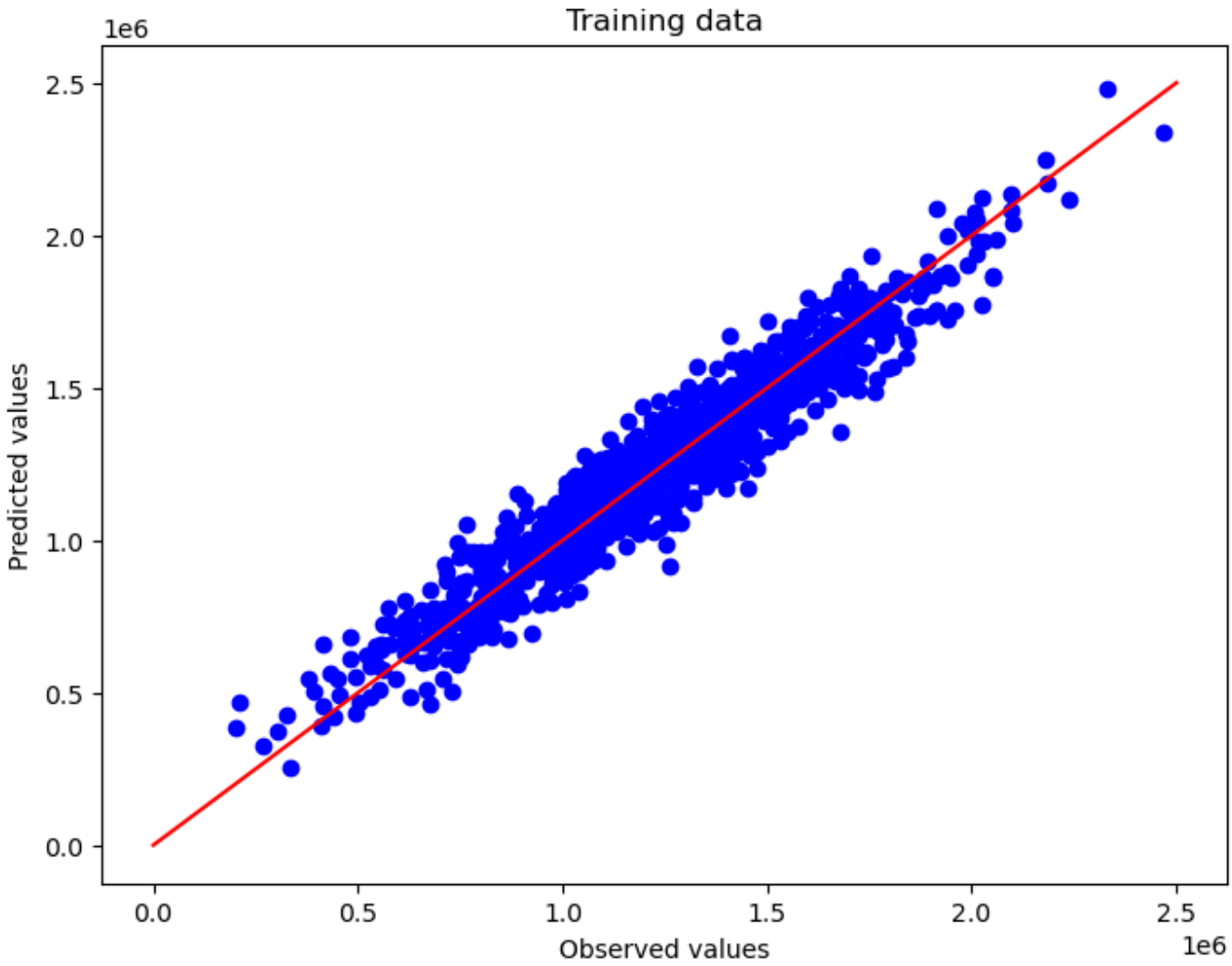
	Coefficient
Income	21.659750
House_Age	164702.184429
Number_Rooms	121003.907115
Population	15.265506

Coefficient de détermination R2

0.9179227041506954

91,8% de la variabilité du prix des maisons sont expliquées par le modele

Prédiction du modèle



By Mariam Sylla

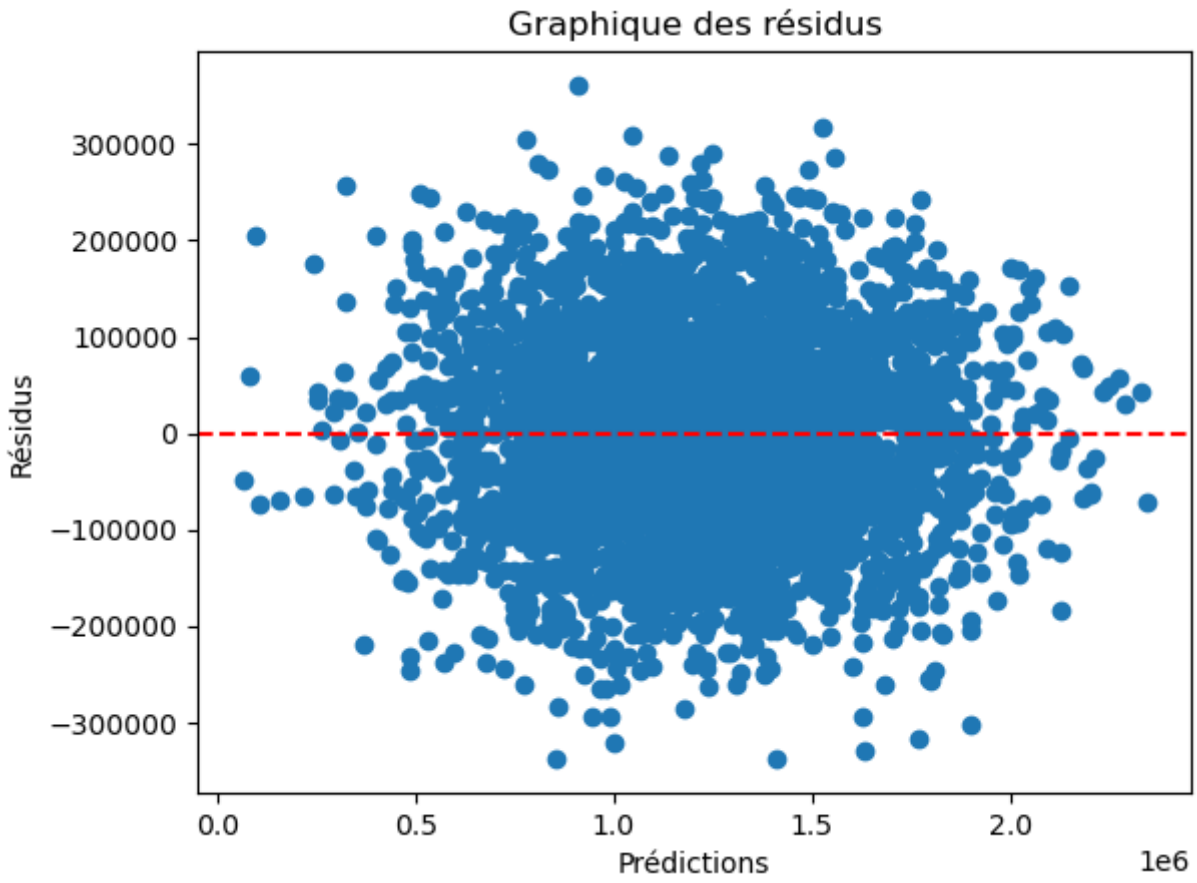
	Valeur Actuelle	Valeur predicte	Difference
1501	1.339096e+06	1.309580e+06	29516.062793
2586	1.251794e+06	1.238948e+06	12846.431398
2653	1.340095e+06	1.247877e+06	92217.690017
1055	1.431508e+06	1.229480e+06	202027.616845
705	1.042374e+06	1.066189e+06	-23815.065951
106	1.555321e+06	1.546651e+06	8669.401081
589	1.250882e+06	1.096610e+06	154271.860801
2468	1.039381e+06	8.313095e+05	208071.231561
2413	8.324752e+05	7.862380e+05	46237.232199
1600	1.420648e+06	1.470839e+06	-50190.912784
2464	6.137883e+05	6.702689e+05	-56480.535429
228	1.702406e+06	1.610862e+06	91543.762158
915	9.135871e+05	1.003612e+06	-90025.302639
794	1.675557e+06	1.800295e+06	-124737.445886
3021	1.279161e+06	1.289322e+06	-10161.267131
3543	9.496844e+05	1.087253e+06	-137568.408999
1073	1.372994e+06	1.420934e+06	-47939.949904
3351	1.148564e+06	1.082556e+06	66007.931715
1744	8.469394e+05	8.065626e+05	40376.865153
1084	1.002193e+06	9.314608e+05	70731.741130

```
          Apprentissage          Test
Métrique
MAE        8.148938e+04  8.085779e+04
MSE        1.026333e+10  1.007372e+10
RMSE       1.013081e+05  1.003679e+05
R2         9.179227e-01  9.181214e-01
```

Vérification des hypothèses du modèle

Homoscédasticité des résidus

By Mariam Sylla



Ici on voit clairement que les residus sont homogène car on ne voit aucune tendance avec les courbes.

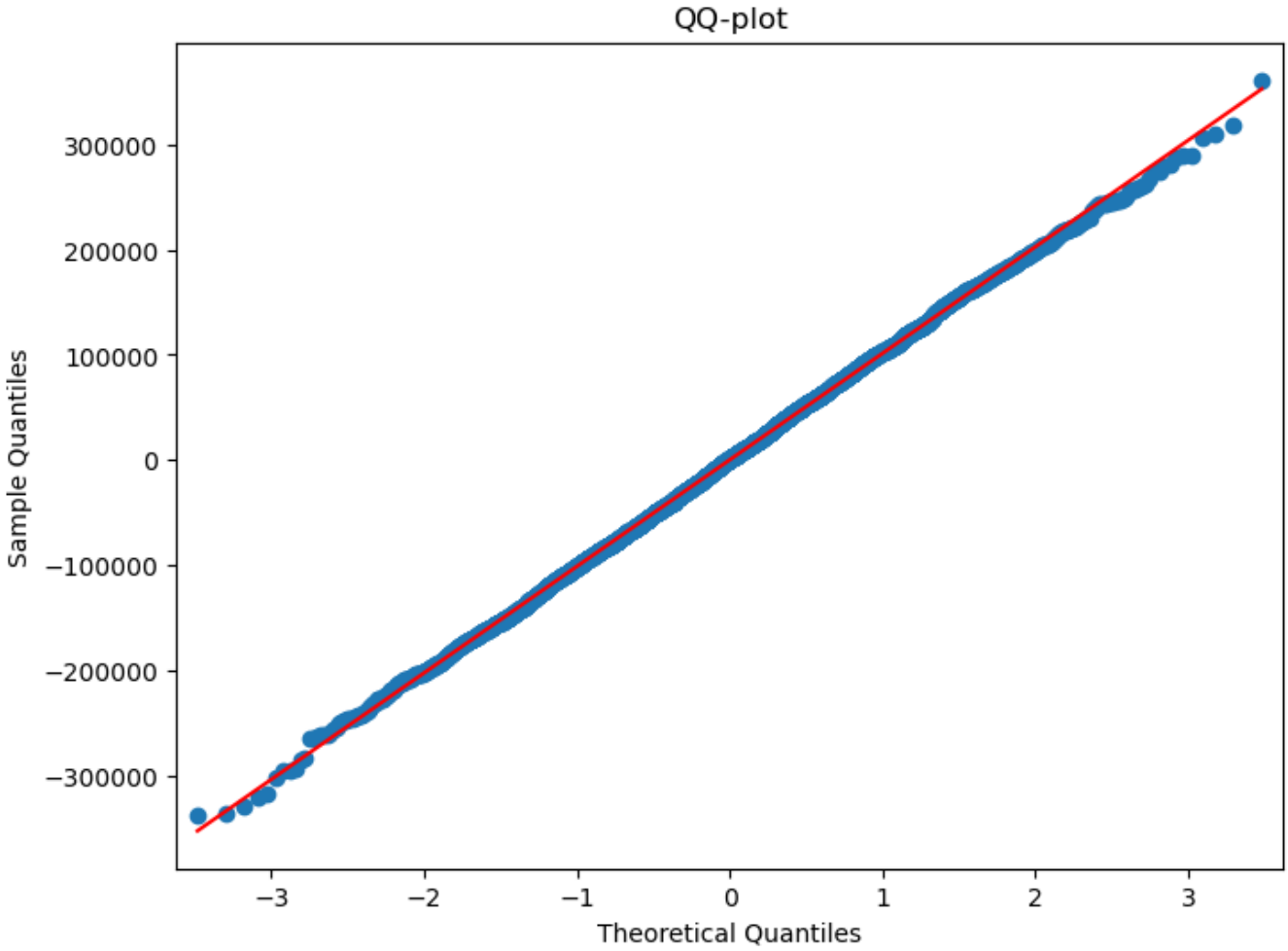
Absence de multi colinéarité entre les variables

	feature	VIF	Multicolinearité?
0	Income	29.632060	Oui
1	House_Age	27.446757	Oui
2	Number_Rooms	32.156187	Oui
3	Population	12.813187	Oui

Les variables ne sont pas correlées entre elles car les VIF sont <5

Normalité des residus

By Mariam Sylla



	Nom du test	Statistique de test	p-valeur	Normalité
0	Shapiro-Wilk	0.999496	0.386034	Oui

La normalité est verifié car p-valeur>5%

Indépendance des erreurs

```
Test de Durbin-Watson : statistique = 1.966666
les residus sont independants
```

La verification de cette hypothese peut omise tant qu'on ne travaille pas sur les series temporelles, car elle est rarement verifié

vérification de la Nullité de l'Esperance des erreurs

```
8.207280188798904e-11
```

Cette hypthosese est toujours verifié

Utilisation du modèle pour réaliser des prédictions

By Mariam Sylla

```
Entrer le revenu de la famille:
>>>79545
Entrez l'age de la maison que vous souhaitez:
>>>6
Entrez le nombre de pièces que vous voulez:
>>>7
Entrez la densité de la population que vous voulez:
>>>23086
le prix de cette maison vaut [1275041.38999503]
```

Régression logistique

Ici nous allons faire une régression logistique pour prédire si un patient a un risque de maladie cardiaque ou pas. Nous avons une base de données qui contient les informations sur le genre(**sexe**), l'age(**age**),si la personne fume ou pas (**currentSmoker**),le nombre de cigarette fumé par jour (**cigsPerDay**),si la personne est diabetique ou pas (**diabetes**), Le taux de cholesterol (**totChol**),la tension arterielle(**sysBP**),pression artérielle diastolique(**diaBP**), l'indice de masse corporelle(**BMI**),la fréquence cardiaque(**heartRate**),le taux de glucose dans le sang(**glucose**) des patients et la variable cible **risque** qui est à prédire, les valeurs de 1 indiquant le risque et de 0 indiquant l'absence risque.

	sexe	age	currentSmoker	cigsPerDay	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	ris
0	female	39	no	0.0	no	195.0	106.0	70.0	26.97	80.0	77.0	
1	male	46	no	0.0	no	250.0	121.0	81.0	28.73	95.0	76.0	
2	female	48	yes	20.0	no	245.0	127.5	80.0	25.34	75.0	70.0	
3	male	61	yes	30.0	no	225.0	150.0	95.0	28.58	65.0	103.0	
4	male	46	yes	23.0	no	285.0	130.0	84.0	23.10	85.0	85.0	
...	...	...	...	...	...	...	...	...	...	...	...	
4233	female	50	yes	1.0	no	313.0	179.0	92.0	25.97	66.0	86.0	
4234	female	51	yes	43.0	no	207.0	126.5	80.0	19.71	65.0	68.0	
4235	male	48	yes	20.0	no	248.0	131.0	72.0	22.00	84.0	86.0	
4236	male	44	yes	15.0	no	210.0	126.5	87.0	19.16	86.0	NaN	
4237	male	52	no	0.0	no	269.0	133.5	83.0	21.47	80.0	107.0	

4238 rows × 12 columns

Nettoyage des données

By Mariam Sylla

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4238 entries, 0 to 4237
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   sexe            4238 non-null   object 
 1   age             4238 non-null   int64  
 2   currentSmoker   4238 non-null   object 
 3   cigsPerDay      4209 non-null   float64
 4   diabetes        4238 non-null   object 
 5   totChol         4188 non-null   float64
 6   sysBP           4238 non-null   float64
 7   diaBP           4238 non-null   float64
 8   BMI             4219 non-null   float64
 9   heartRate       4237 non-null   float64
 10  glucose         3850 non-null   float64
 11  risque          4238 non-null   int64  
dtypes: float64(7), int64(2), object(3)
memory usage: 397.4+ KB
```

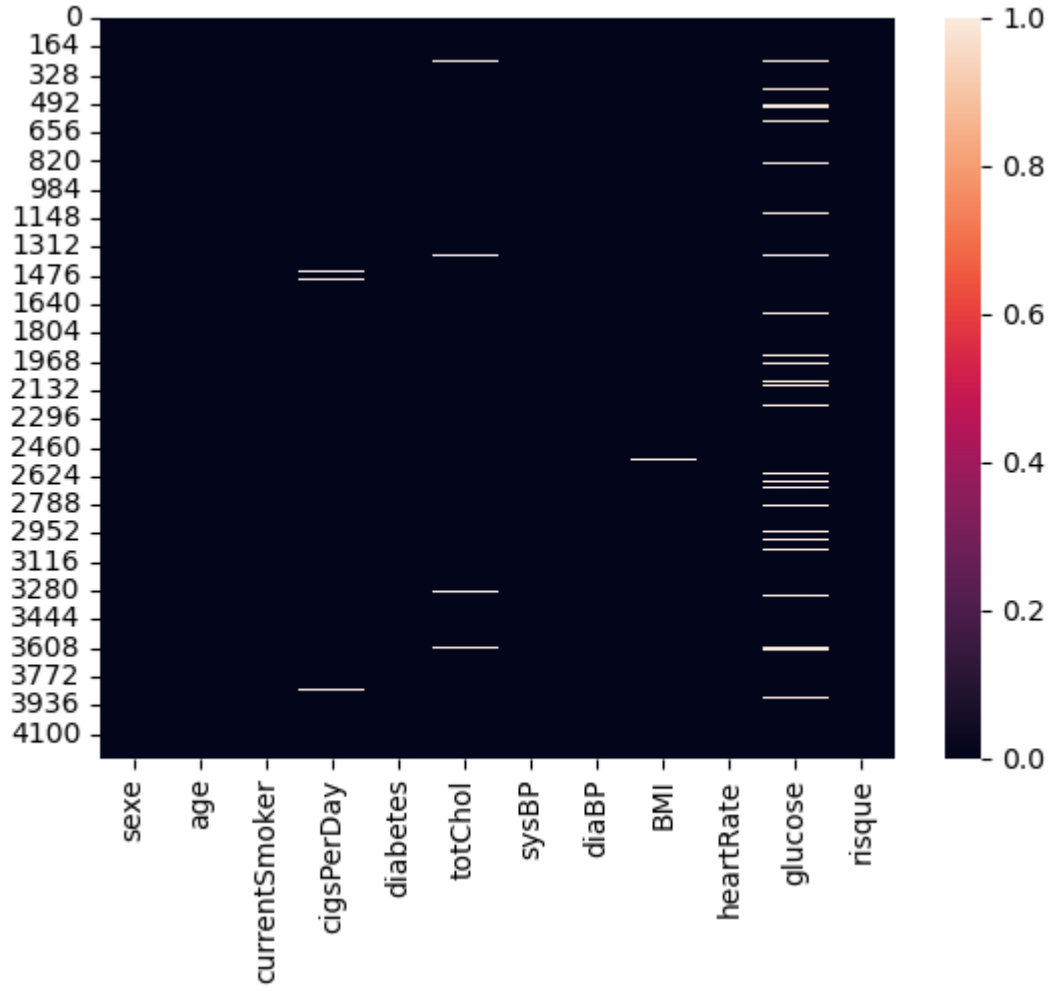
	count	mean	std	min	25%	50%	75%	max
age	4238.0	49.584946	8.572160	32.00	42.00	49.0	56.000	70.0
cigsPerDay	4209.0	9.003089	11.920094	0.00	0.00	0.0	20.000	70.0
totChol	4188.0	236.721585	44.590334	107.00	206.00	234.0	263.000	696.0
sysBP	4238.0	132.352407	22.038097	83.50	117.00	128.0	144.000	295.0
diaBP	4238.0	82.893464	11.910850	48.00	75.00	82.0	89.875	142.5
BMI	4219.0	25.802008	4.080111	15.54	23.07	25.4	28.040	56.8
heartRate	4237.0	75.878924	12.026596	44.00	68.00	75.0	83.000	143.0
glucose	3850.0	81.966753	23.959998	40.00	71.00	78.0	87.000	394.0
risque	4238.0	0.151958	0.359023	0.00	0.00	0.0	0.000	1.0

Detection des valeurs manquantes

By Mariam Sylla

	Nombre de valeurs manquantes	Proportion de valeurs manquantes
glucose	388	0.091553
totChol	50	0.011798
cigsPerDay	29	0.006843
BMI	19	0.004483
heartRate	1	0.000236
sexe	0	0.000000
age	0	0.000000
currentSmoker	0	0.000000
diabetes	0	0.000000
sysBP	0	0.000000
diaBP	0	0.000000
risque	0	0.000000

```
<Axes: >
```



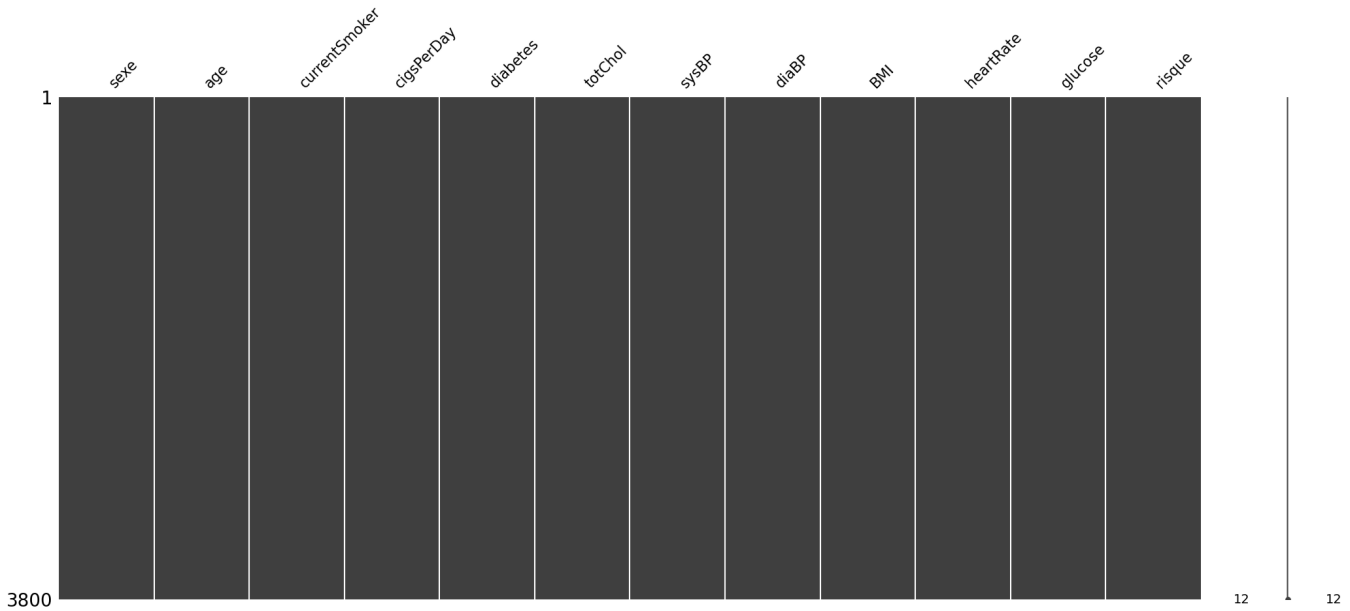
Nous avons des valeurs manquantes, comme la proportion n'est pas elevé nous allons les supprimer

By Mariam Sylla

	sexe	age	currentSmoker	cigsPerDay	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	ris
14	male	39	yes	9.0	no	226.0	114.0	64.0	22.35	85.0	NaN	
21	male	43	no	0.0	no	185.0	123.5	77.5	29.89	70.0	NaN	
26	male	60	no	0.0	no	260.0	110.0	72.5	26.59	65.0	NaN	
42	male	52	no	0.0	no	NaN	148.0	92.0	25.09	70.0	NaN	
54	male	39	yes	20.0	no	209.0	115.0	75.0	22.54	90.0	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...
4185	female	58	no	0.0	no	NaN	116.5	71.0	27.04	70.0	86.0	
4208	male	51	yes	9.0	no	340.0	152.0	76.0	25.74	70.0	NaN	
4229	male	51	yes	20.0	no	251.0	140.0	80.0	25.60	75.0	NaN	
4230	male	56	yes	3.0	no	268.0	170.0	102.0	22.89	57.0	NaN	
4236	male	44	yes	15.0	no	210.0	126.5	87.0	19.16	86.0	NaN	

438 rows × 12 columns

Suppression des valeurs manquantes



Detection des doublons

```
0
```

Detection des Outliers

By Mariam Sylla

```
--------------------------- age -------------------------
No Outliers Present


--------------------------- cigsPerDay -------------------------
        sexe  age currentSmoker  cigsPerDay diabetes  totChol  sysBP  diaBP  \
327   female   56            yes        60.0       no    246.0  125.0   79.0
721   female   59            yes        60.0       no    298.0  153.5  105.0
1054  female   58            yes        60.0       no    250.0  150.0   97.0
1452  female   39            yes        60.0       no    215.0  112.0   65.0
1468  female   50            yes        60.0       no    340.0  134.0   95.0
1488  female   37            yes        60.0       no    254.0  122.5   82.5
1849  female   48            yes        60.0       no    252.0  104.0   73.5
2709  female   46            yes        60.0       no    285.0  121.0   82.0
3008  female   40            yes        70.0       no    210.0  132.0   86.0
3673  female   48            yes        60.0       no    232.0  136.0   81.0
3928  female   67            yes        60.0       no    261.0  170.0  100.0

        BMI  heartRate  glucose  risque
327   29.64       70.0     85.0       0
721   25.05       70.0     84.0       0
1054  32.00       75.0     65.0       0
1452  23.60       59.0     78.0       0
1468  30.46       85.0     86.0       0
1488  23.87       88.0     83.0       0
1849  23.03       70.0     77.0       0
2709  27.62       70.0     79.0       0
3008  31.57       98.0     80.0       0
3673  25.83       80.0     78.0       0
3928  22.71       72.0     79.0       1


--------------------------- totChol -------------------------
        sexe  age currentSmoker  cigsPerDay diabetes  totChol  sysBP  diaBP  \
194     male   42            no         0.0       no    464.0  128.0   87.0
259     male   60            no         0.0       no    352.0  197.5  105.0
333     male   55            no         0.0       no    368.0  204.0   94.0
533     male   53            no         0.0       no    370.0  123.0   83.0
543   female   47           yes        18.0       no    439.0  145.0   74.0
617     male   51           yes         3.0       no    398.0  161.0   96.0
670     male   65            no         0.0       no    355.0  138.0   79.0
920     male   60           yes        15.0       no    353.0  116.0   82.0
926   female   61           yes        20.0       no    360.0  157.0   99.0
998     male   57            no         0.0       no    372.0  122.0   80.0
1103    male   60           yes        20.0       no    352.0  149.0   73.0
1111    male   52            no         0.0      yes    600.0  159.5   94.0
1150    male   46           yes        10.0       no    392.0  113.0   68.0
1389    male   51            no         0.0       no    358.0  134.0   87.0
1447    male   60            no         0.0       no    391.0  114.0   64.0
1544  female   42           yes        20.0       no    410.0  116.0   83.0
1691    male   64            no         0.0       no    372.0  169.0   85.0
1710    male   57            no         0.0       no    366.0  146.5   80.0
1937    male   59            no         0.0       no    410.0  142.0   79.0
2024    male   62            no         0.0      yes    390.0  184.5   83.0
2206  female   46           yes        15.0       no    405.0  181.5  102.5
2208    male   42           yes        10.0       no    359.0  115.0   71.0
2349    male   63            no         0.0       no    380.0  175.0   78.0
2363  female   42           yes         5.0       no    355.0  113.0   81.0
2368    male   54           yes         5.0       no    390.0  150.0   94.0
2488    male   67           yes        15.0       no    371.0  166.0   85.0
2525  female   38           yes        20.0       no    113.0  120.0   83.5
2602    male   60            no         0.0       no    354.0  130.0   82.5
2607    male   57            no         0.0       no    382.0  133.0   77.0
```

By Mariam Sylla

2671	male	59	no	0.0	no	364.0	142.0	84.0
2797	female	43	no	0.0	no	367.0	141.0	82.5
2972	male	51	yes	15.0	no	352.0	136.5	87.0
2985	male	57	no	0.0	no	432.0	153.0	85.0
3160	female	51	yes	9.0	no	696.0	157.0	87.0
3165	female	44	yes	30.0	no	363.0	140.0	87.0
3394	male	57	yes	9.0	no	382.0	140.0	94.0
3418	male	63	no	0.0	no	361.0	167.0	100.0
3474	female	42	yes	15.0	no	453.0	158.0	108.0
3532	female	44	yes	3.0	no	352.0	164.0	119.0
3806	male	52	yes	20.0	no	410.0	105.0	67.5
3816	male	56	no	0.0	no	391.0	126.0	84.0
3844	male	62	yes	20.0	yes	358.0	215.0	110.0
3916	female	62	yes	30.0	no	373.0	138.5	85.0
4001	male	58	no	0.0	no	385.0	165.0	95.0

	BMI	heartRate	glucose	risque
194	22.90	72.0	72.0	1
259	36.29	75.0	95.0	1
333	25.20	100.0	81.0	0
533	24.64	63.0	74.0	1
543	22.42	100.0	90.0	1
617	23.63	77.0	83.0	0
670	28.38	75.0	108.0	0
920	22.66	85.0	71.0	0
926	28.74	95.0	73.0	0
998	21.02	65.0	81.0	0
1103	25.96	80.0	79.0	0
1111	28.27	78.0	140.0	1
1150	23.35	70.0	63.0	0
1389	29.36	75.0	87.0	1
1447	24.57	82.0	83.0	0
1544	21.68	90.0	83.0	0
1691	26.01	75.0	79.0	1
1710	24.19	85.0	73.0	0
1937	25.58	78.0	90.0	0
2024	18.99	87.0	47.0	0
2206	26.33	98.0	97.0	1
2208	24.46	75.0	68.0	0
2349	20.15	68.0	95.0	1
2363	26.17	90.0	71.0	0
2368	27.34	75.0	71.0	0
2488	25.35	100.0	86.0	0
2525	30.34	78.0	85.0	0
2602	26.76	65.0	79.0	0
2607	24.27	75.0	81.0	0
2671	26.24	67.0	70.0	0
2797	25.62	92.0	90.0	0
2972	25.79	73.0	67.0	0
2985	26.13	98.0	75.0	1
3160	24.44	95.0	84.0	0
3165	26.44	95.0	79.0	0
3394	21.20	98.0	70.0	0
3418	27.31	85.0	103.0	1
3474	28.89	90.0	110.0	0
3532	28.92	73.0	72.0	1
3806	27.33	75.0	90.0	0
3816	24.83	80.0	78.0	0
3844	37.62	110.0	368.0	1
3916	23.35	80.0	67.0	0
4001	41.66	82.0	91.0	0

---------- sysBP ----------

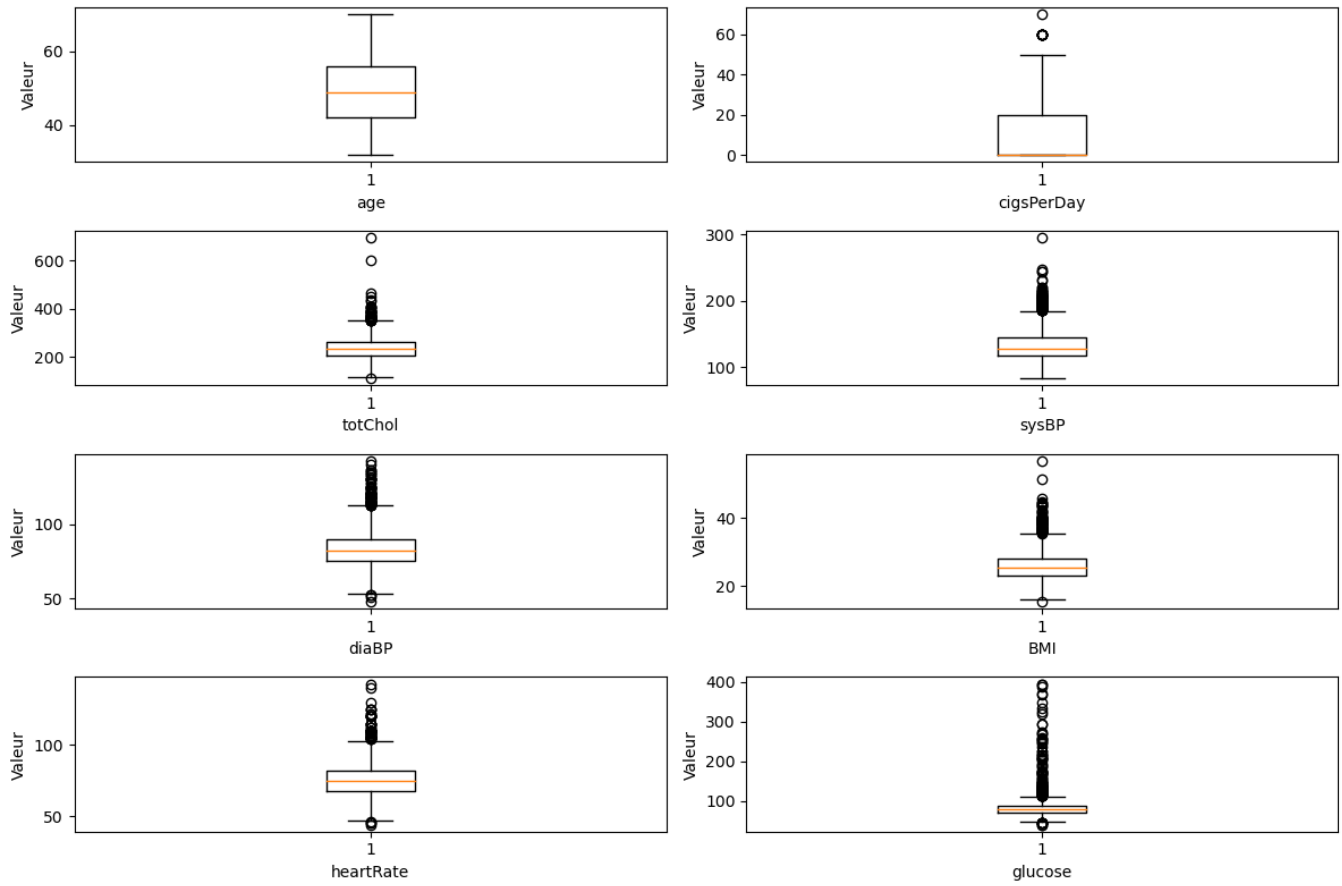
```
        sexe  age currentSmoker  cigsPerDay diabetes  totChol  sysBP  diaBP  \
44      male   53            no         0.0      yes    311.0  206.0   92.0
66      male   62            no         0.0      yes    212.0  190.0   99.0
87      male   61           yes         1.0       no    326.0  200.0  104.0
108     male   66            no         0.0       no    278.0  187.0   88.0
153     male   66            no         0.0       no    214.0  212.0  104.0
...      ...  ...           ...         ...      ...      ...    ...    ...
4123  female   51            no         0.0       no    268.0  206.0  116.0
4173    male   54            no         0.0       no    302.0  210.0  127.5
4193    male   63            no         0.0       no    306.0  195.0  105.0
4222  female   53            no         0.0       no    289.0  188.0  110.0
4228    male   50            no         0.0      yes    260.0  190.0  130.0

        BMI  heartRate  glucose  risque
44    21.51       76.0    215.0       1
66    29.64      100.0    202.0       0
87    38.46       57.0     78.0       0
108   40.52       90.0     84.0       1
153   25.32       57.0     84.0       1
...     ...        ...      ...     ...
4123  26.35       98.0     70.0       1
4173  31.98       68.0     79.0       0
4193  27.96       75.0     87.0       1
4222  26.70       70.0     63.0       0
4228  43.67       85.0    260.0       0

[113 rows x 12 columns]


--------------------------- diaBP ---------------------------
        sexe  age currentSmoker  cigsPerDay diabetes  totChol  sysBP  diaBP  \
28      male   61            no         0.0       no    272.0  182.0  121.0
46      male   65            no         0.0       no    252.0  179.5  114.0
158     male   49            no         0.0       no    254.0  191.0  124.5
249     male   60           yes        20.0      yes    180.0  200.0  122.5
409   female   44           yes        10.0       no    229.0  177.5  120.0
...      ...  ...           ...         ...      ...      ...    ...    ...
4075  female   63           yes        25.0       no    203.0  192.5  125.0
4076    male   61            no         0.0      yes    265.0  200.0  125.0
4123  female   51            no         0.0       no    268.0  206.0  116.0
4173    male   54            no         0.0       no    302.0  210.0  127.5
4228    male   50            no         0.0      yes    260.0  190.0  130.0

        BMI  heartRate  glucose  risque
28    32.80       85.0     65.0       1
46    30.47       90.0     87.0       0
158   28.35       78.0     54.0       0
249   44.27       88.0    150.0       0
409   39.88      104.0     78.0       0
...     ...        ...      ...     ...
4075  26.18       80.0     83.0       1
4076  29.50       68.0    256.0       1
4123  26.35       98.0     70.0       1
4173  31.98       68.0     79.0       0
4228  43.67       85.0    260.0       0

[71 rows x 12 columns]


--------------------------- BMI ---------------------------
        sexe  age currentSmoker  cigsPerDay diabetes  totChol  sysBP  diaBP  \
35    female   37            no         0.0       no    225.0  124.5   92.5
37    female   52            no         0.0      yes    178.0  160.0   98.0
78      male   45            no         0.0       no    183.0  151.0  101.0
```

By Mariam Sylla

```
87      male   61           yes         1.0       no    326.0  200.0  104.0
108     male   66            no         0.0       no    278.0  187.0   88.0
...      ...  ...           ...         ...      ...      ...    ...    ...
4001    male   58            no         0.0       no    385.0  165.0   95.0
4132    male   57            no         0.0       no    259.0  170.0  101.0
4190    male   41            no         0.0       no    229.0  150.0   89.0
4215    male   63            no         0.0      yes    236.0  155.0   82.0
4228    male   50            no         0.0      yes    260.0  190.0  130.0

        BMI  heartRate  glucose  risque
35    38.53       95.0     83.0       0
37    40.11       75.0    225.0       0
78    45.80       80.0     63.0       0
87    38.46       57.0     78.0       0
108   40.52       90.0     84.0       1
...     ...        ...      ...     ...
4001  41.66       82.0     91.0       0
4132  38.17       85.0     75.0       0
4190  36.07       75.0     92.0       0
4215  39.17       78.0     79.0       0
4228  43.67       85.0    260.0       0

[88 rows x 12 columns]


--------------------------- heartRate -------------------------
        sexe  age currentSmoker  cigsPerDay diabetes  totChol  sysBP  diaBP  \
162     male   47            no         0.0       no    174.0  118.0   86.5
270     male   54            no         0.0       no    273.0  139.0   98.0
339     male   64            no         0.0       no    312.0  160.0   82.0
358     male   40           yes        20.0       no    210.0  118.0   79.0
409   female   44           yes        10.0       no    229.0  177.5  120.0
...      ...  ...           ...         ...      ...      ...    ...    ...
3964    male   39            no         0.0       no    213.0  125.0   87.0
4053    male   44            no         0.0       no    160.0  107.0   69.0
4070    male   40            no         0.0       no    202.0  158.0  103.0
4164  female   39           yes        20.0       no    287.0  136.0   86.0
4195    male   40           yes         9.0       no    207.0  124.0   78.0

        BMI  heartRate  glucose  risque
162   26.15      110.0     86.0       0
270   29.06      110.0     73.0       1
339   27.59      140.0     94.0       0
358   21.21      130.0     84.0       0
409   39.88      104.0     78.0       0
...     ...        ...      ...     ...
3964  16.73      110.0     75.0       0
4053  18.63      125.0     78.0       0
4070  28.35      125.0     80.0       0
4164  19.00      112.0     83.0       0
4195  22.90       46.0     66.0       0

[81 rows x 12 columns]


--------------------------- glucose -------------------------
        sexe  age currentSmoker  cigsPerDay diabetes  totChol  sysBP  diaBP  \
22      male   52            no         0.0       no    234.0  148.0   78.0
37    female   52            no         0.0      yes    178.0  160.0   98.0
44      male   53            no         0.0      yes    311.0  206.0   92.0
65      male   63            no         0.0       no    252.0  154.0   87.0
66      male   62            no         0.0      yes    212.0  190.0   99.0
...      ...  ...           ...         ...      ...      ...    ...    ...
4115    male   63            no         0.0       no    250.0  190.0   88.0
```

By Mariam Sylla

```
4118    male   37            no         0.0       no    160.0  137.0   82.0
4203  female   63           yes        10.0      yes    240.0  146.0   84.0
4209  female   65            no         0.0       no    286.0  135.0   80.0
4228    male   50            no         0.0      yes    260.0  190.0  130.0

        BMI  heartRate  glucose  risque
22    34.17       70.0    113.0       0
37    40.11       75.0    225.0       0
44    21.51       76.0    215.0       1
65    28.60       72.0     45.0       0
66    29.64      100.0    202.0       0
...     ...        ...      ...     ...
4115  24.16       94.0    118.0       1
4118  21.03       94.0    113.0       0
4203  30.48       75.0    120.0       0
4209  28.06       70.0    116.0       0
4228  43.67       85.0    260.0       0

[184 rows x 12 columns]



Variable: age
Nombre de valeurs aberrantes: 19
Proportion de valeurs aberrantes: 0.50%
------------------------------
Variable: cigsPerDay
Nombre de valeurs aberrantes: 11
Proportion de valeurs aberrantes: 0.29%
------------------------------
Variable: totChol
Nombre de valeurs aberrantes: 44
Proportion de valeurs aberrantes: 1.16%
------------------------------
Variable: sysBP
Nombre de valeurs aberrantes: 113
Proportion de valeurs aberrantes: 2.97%
------------------------------
Variable: diaBP
Nombre de valeurs aberrantes: 71
Proportion de valeurs aberrantes: 1.87%
------------------------------
Variable: BMI
Nombre de valeurs aberrantes: 88
Proportion de valeurs aberrantes: 2.32%
------------------------------
Variable: heartRate
Nombre de valeurs aberrantes: 81
Proportion de valeurs aberrantes: 2.13%
------------------------------
Variable: glucose
Nombre de valeurs aberrantes: 184
Proportion de valeurs aberrantes: 4.84%
------------------------------
```



Nous allons supprimer les valeurs aberrante pour toutes les variables independantes

Suppression des valeurs aberrantes

	sexe	age	currentSmoker	cigsPerDay	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	ris
0	female	39	no	0.0	no	195.0	106.0	70.0	26.97	80.0	77.0	
1	male	46	no	0.0	no	250.0	121.0	81.0	28.73	95.0	76.0	
2	female	48	yes	20.0	no	245.0	127.5	80.0	25.34	75.0	70.0	
3	male	61	yes	30.0	no	225.0	150.0	95.0	28.58	65.0	103.0	
4	male	46	yes	23.0	no	285.0	130.0	84.0	23.10	85.0	85.0	
...	...	...	...	...	...	...	...	...	...	...	...	
4232	female	68	no	0.0	no	176.0	168.0	97.0	23.14	60.0	79.0	
4233	female	50	yes	1.0	no	313.0	179.0	92.0	25.97	66.0	86.0	
4234	female	51	yes	43.0	no	207.0	126.5	80.0	19.71	65.0	68.0	
4235	male	48	yes	20.0	no	248.0	131.0	72.0	22.00	84.0	86.0	
4237	male	52	no	0.0	no	269.0	133.5	83.0	21.47	80.0	107.0	

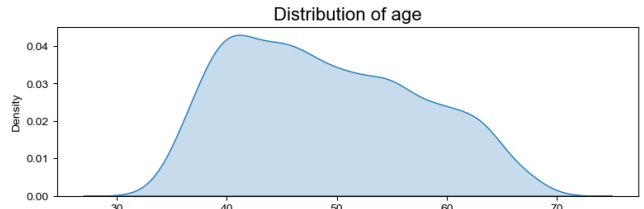
3301 rows × 12 columns

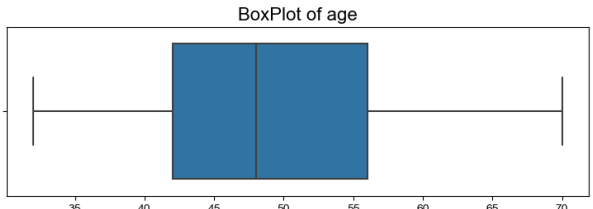
Descripition des données

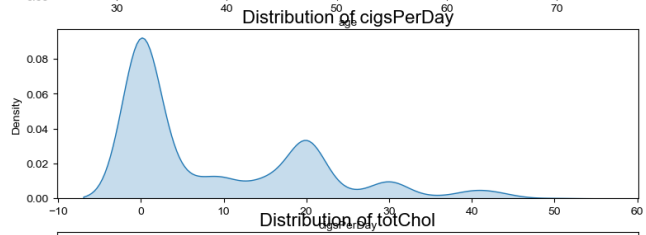
By Mariam Sylla

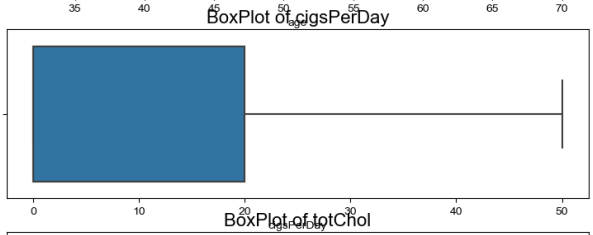
Variables quantitatives

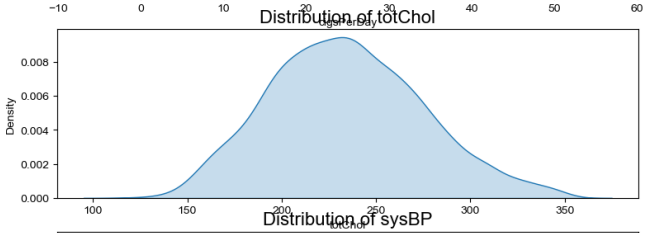
Distribution et Boxplot

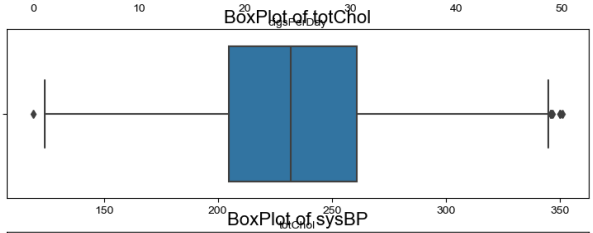


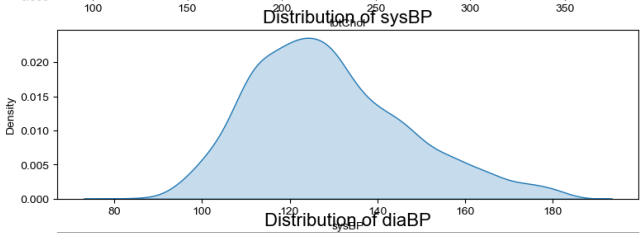


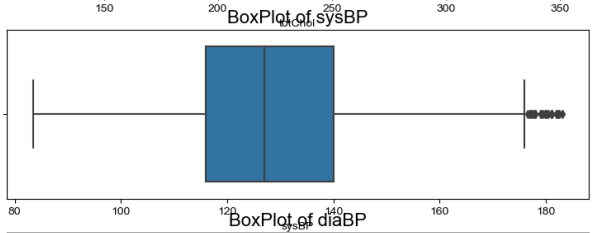


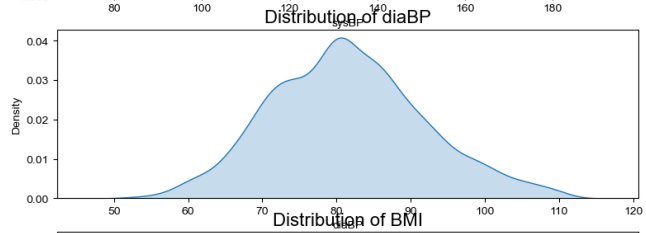


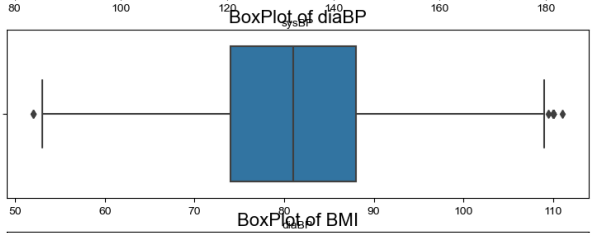


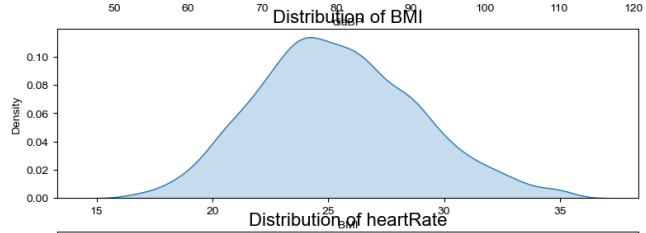


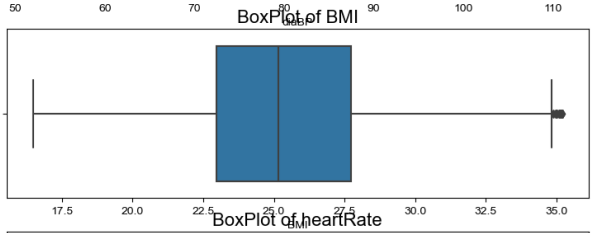


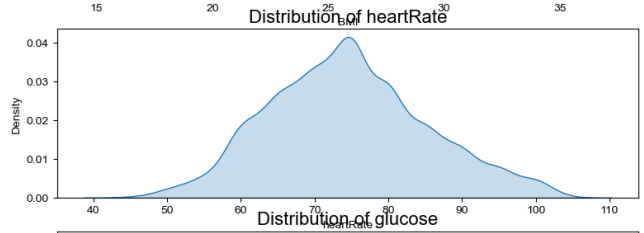


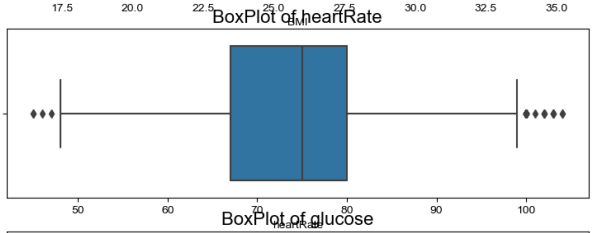


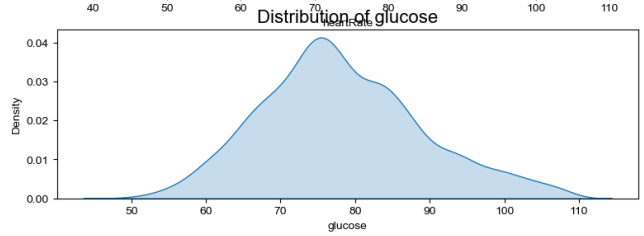


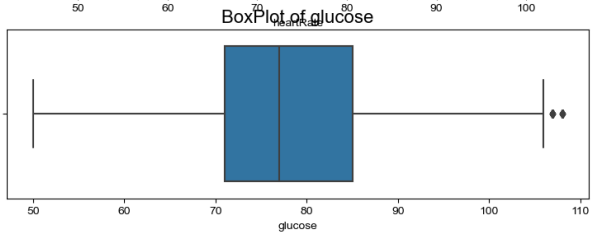




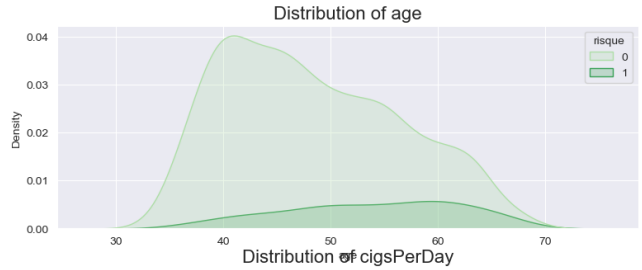


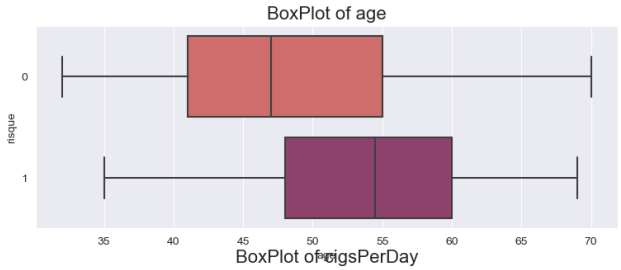


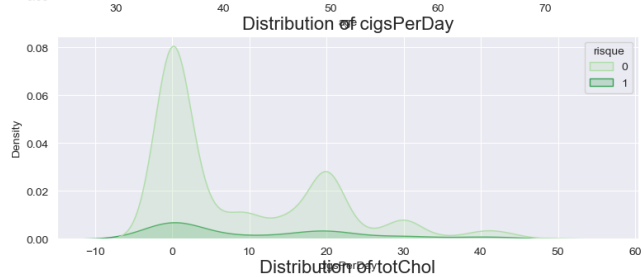


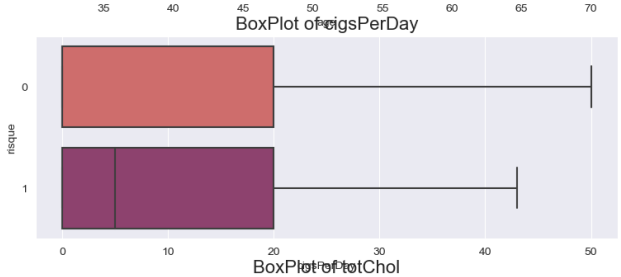


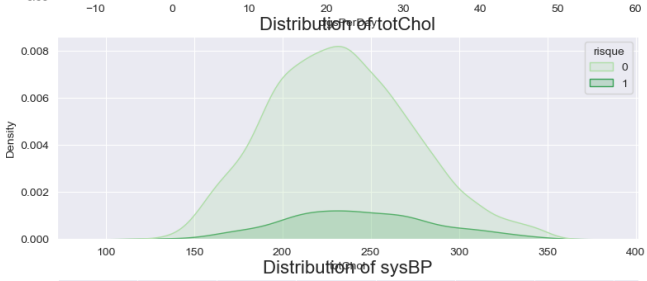
By Mariam Sylla

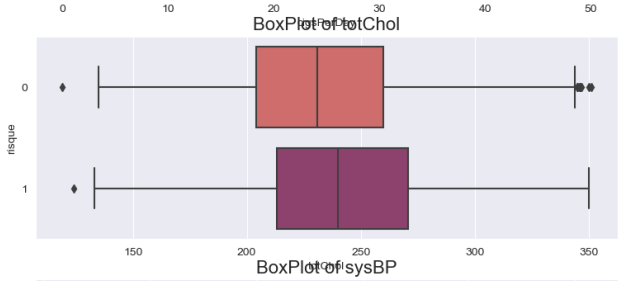


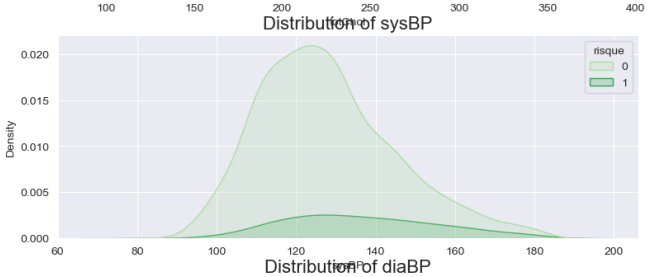


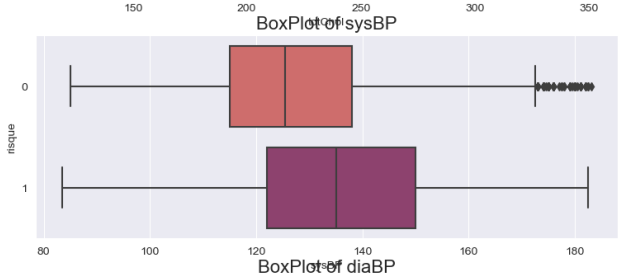


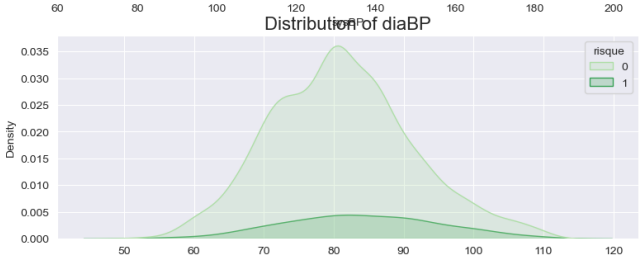


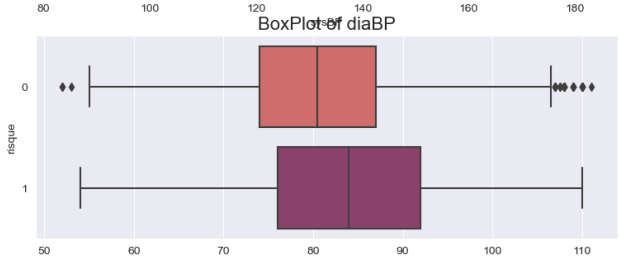


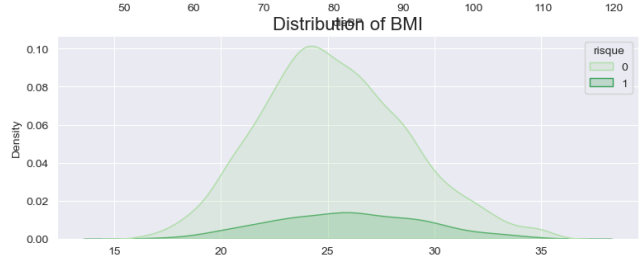


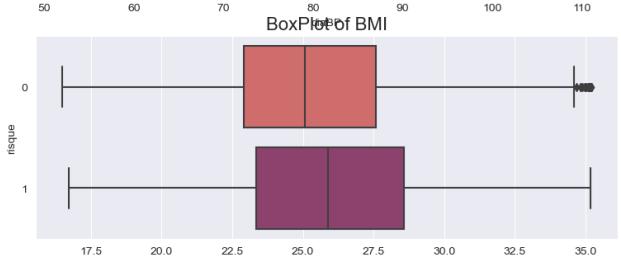


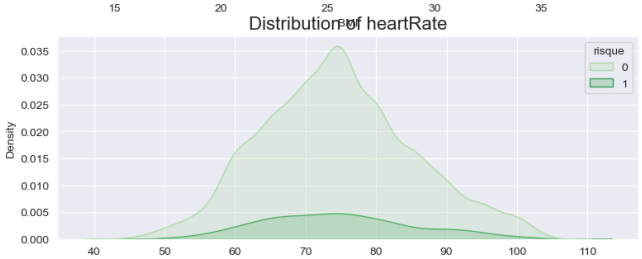


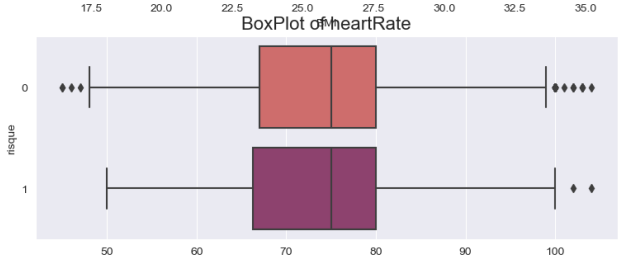


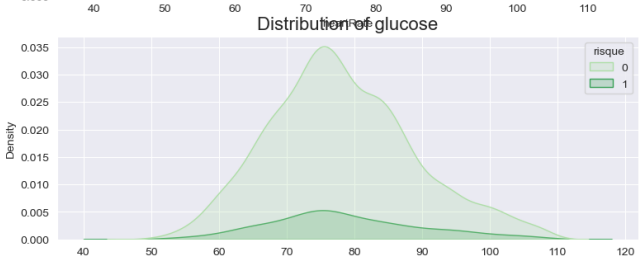


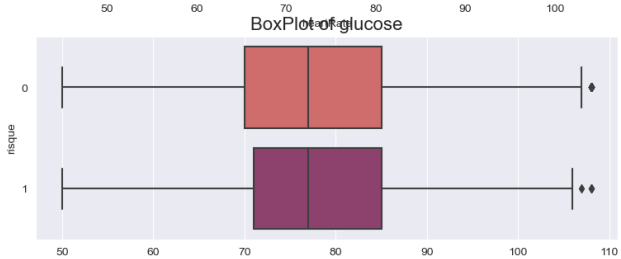












By Mariam Sylla

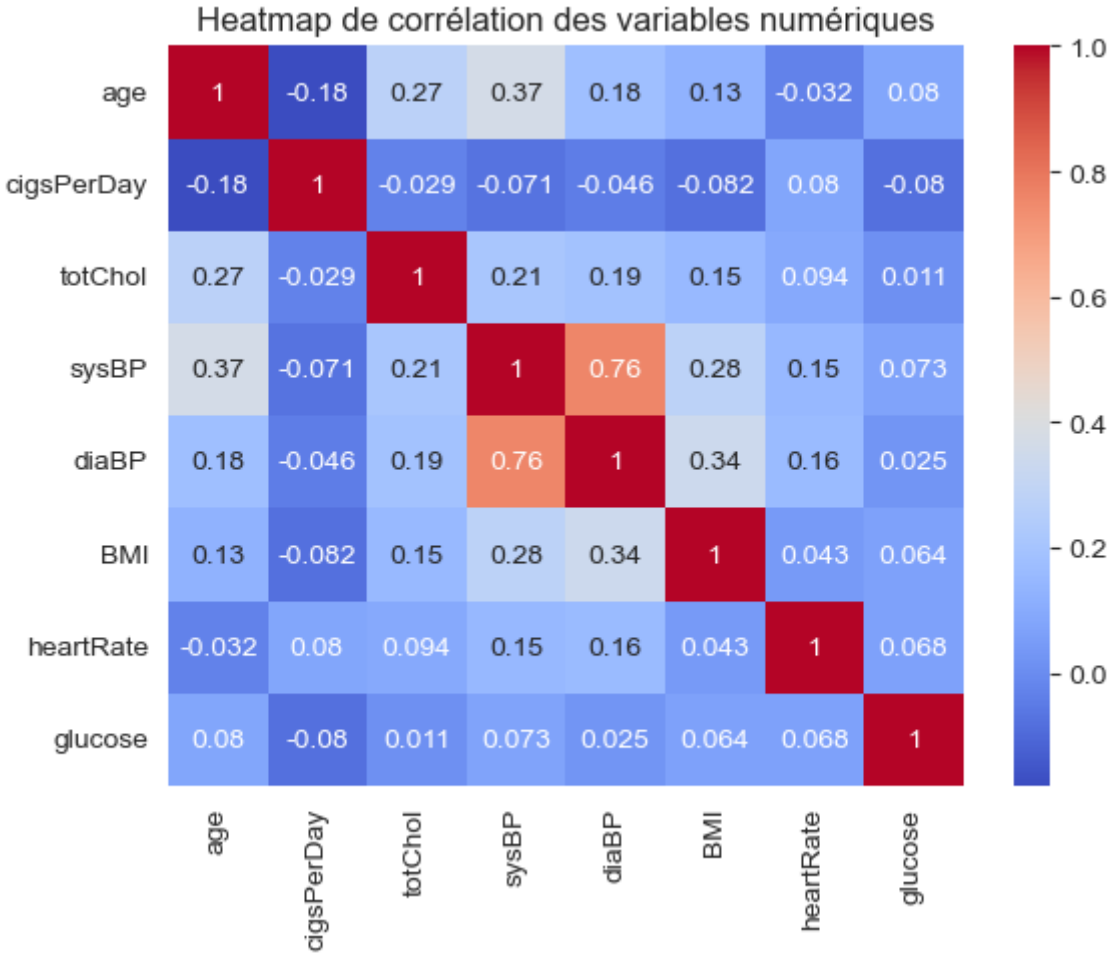
Pour toutes les variables on constate qu'elles dépendent du risque de maladie cardiaque ou non, mais nous allons confirmer cette hypothèse par un test. Nous allons utiliser le test de krukal-wallis car on n'a pas vérifié si les données sont normalement distribuées.

Test Kruskal-wallis afin de confirmer les hypothèses

Variable	P-value	Significative
age	4.525546e-35	Oui
cigsPerDay	4.507093e-04	Oui
totChol	3.405462e-06	Oui
sysBP	8.975951e-20	Oui
diaBP	9.207995e-08	Oui
BMI	1.428031e-04	Oui
heartRate	9.039822e-01	Non
glucose	6.137695e-01	Non

Toutes les variables verifient l'hypothèse à part les variables "heartrate" et "glucose"

Heatmap de corrélation

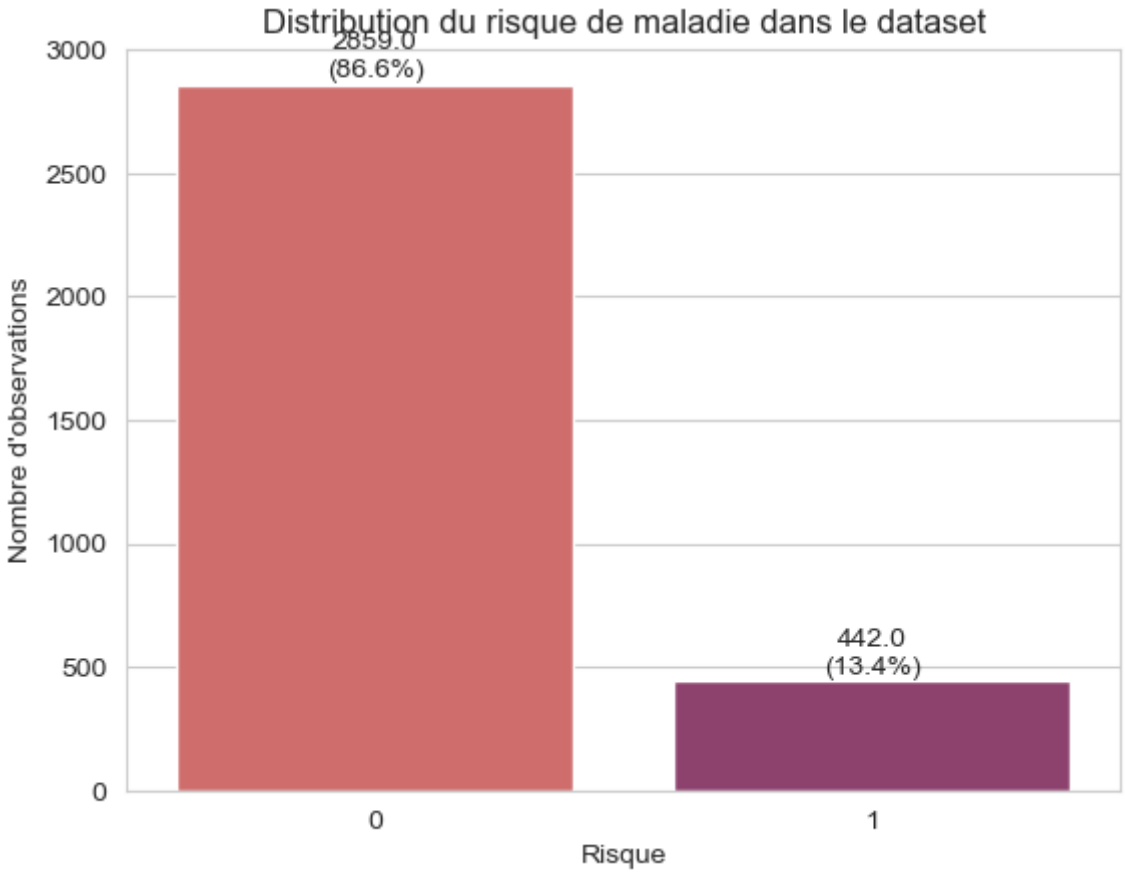


On constate qu'il n'y a pas de forte corrélation entre les variables.

Variables categorielles

Distribution de la variable cible

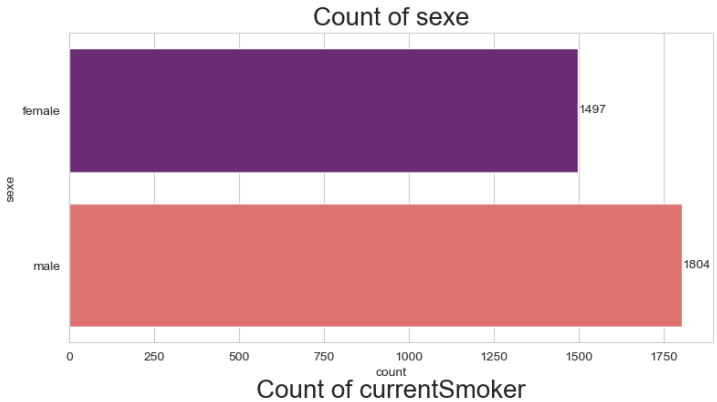
By Mariam Sylla

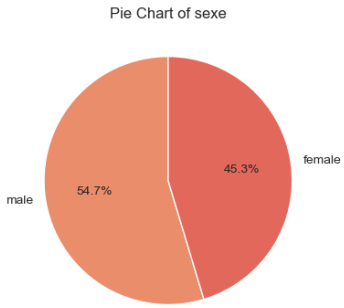


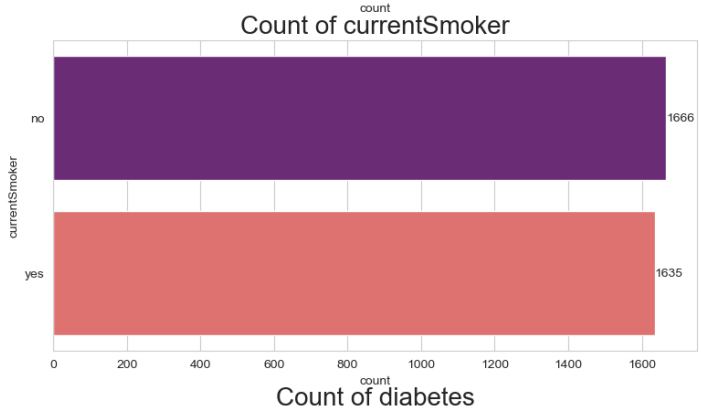
Les deux classes dans la base de données ne sont pas équilibrées, nous avons 87% de personnes qui ne présentent pas de risque de maladie cardiaque contre 13% qui ont un risque. Nous allons utiliser une méthode d'échantillonnage pour équilibrer les deux classes du dataset

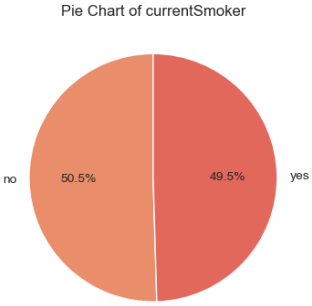
Distribution des variables catégorielles

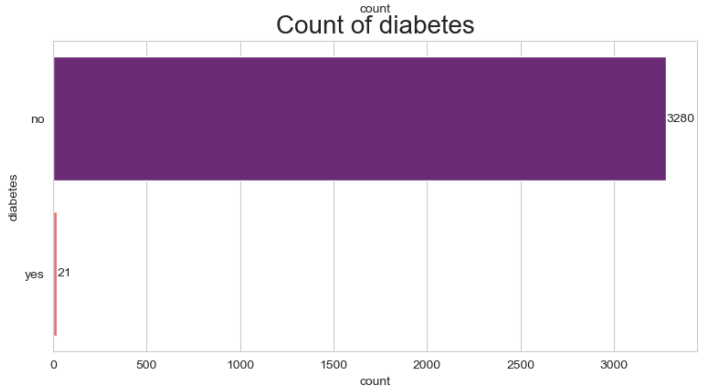
By Mariam Sylla

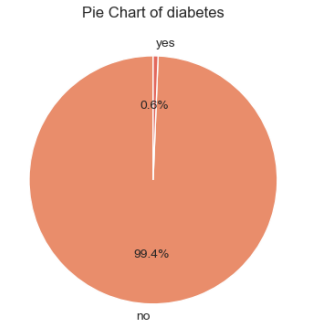




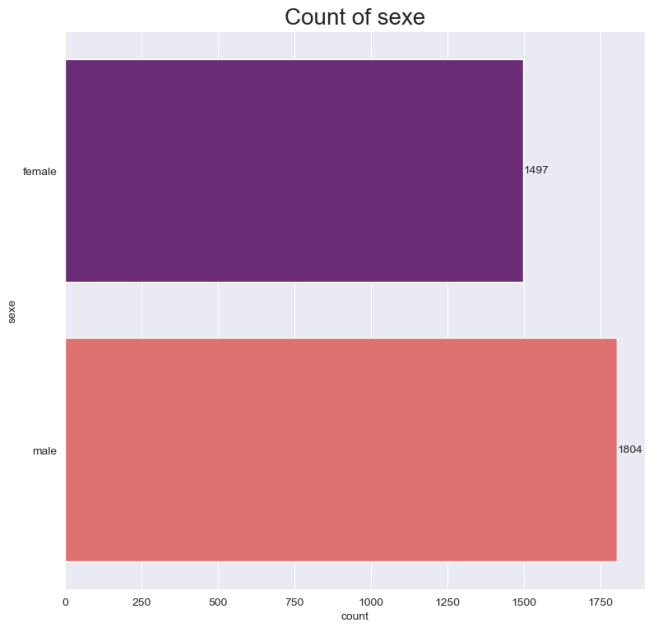


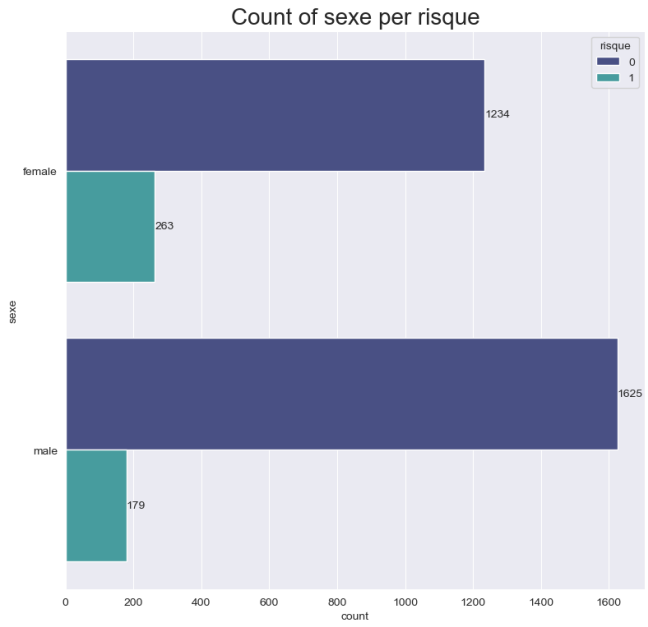


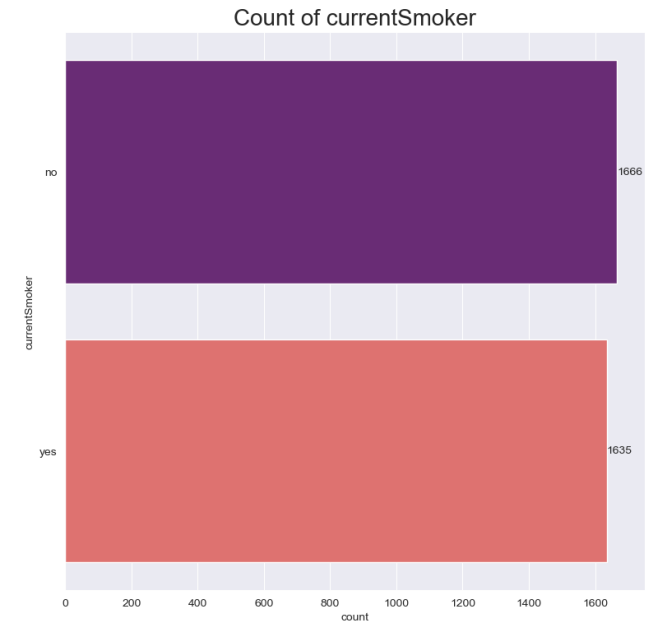


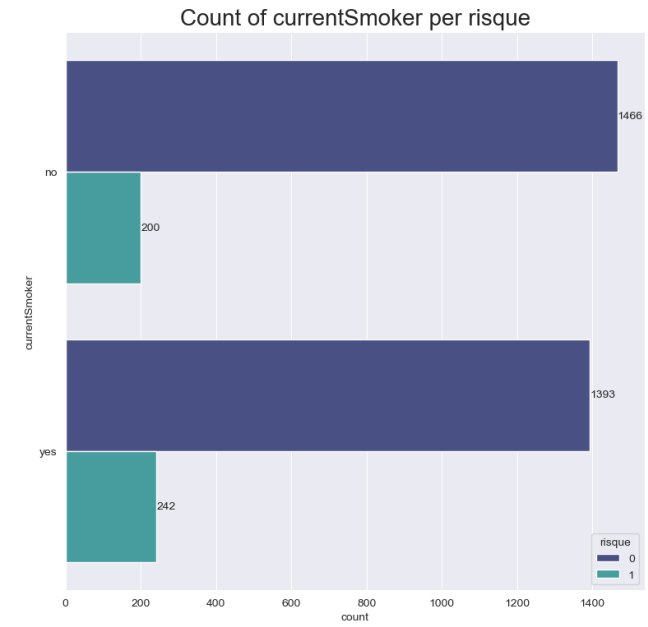


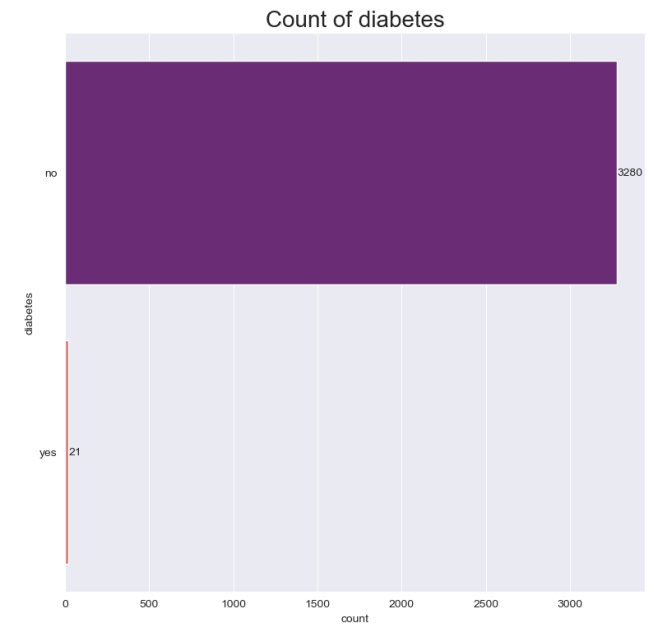
By Mariam Sylla

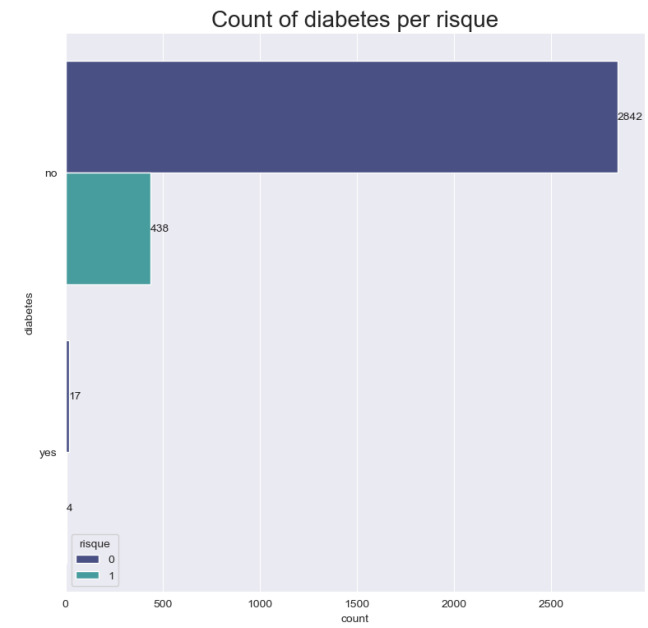












By Mariam Sylla

On constate qu'il existe une relation entre les variables catégorielles et le fait d'avoir ou non un risque de maladie cardiaque, nous allons confirmer cette hypothèse par un test de khi2

Test de Chi2 pour confirmer les hypothèses

```
         Variable       Chi2      P-valeur associations  V de Cramer
0           sexe  40.585784  1.881707e-10          Oui     0.109524
1  currentSmoker   5.325735  2.101275e-02          Oui     0.036204
2       diabetes   0.195678  6.582325e-01          Non     0.000000
```

L'hypothèses est confirmées à part pour la variable "diabetes"

Encodage

	age	cigsPerDay	totChol	sysBP	diaBP	BMI	heartRate	glucose	male	yes	yes
0	39	0.0	195.0	106.0	70.0	26.97	80.0	77.0	0	0	0
1	46	0.0	250.0	121.0	81.0	28.73	95.0	76.0	1	0	0
2	48	20.0	245.0	127.5	80.0	25.34	75.0	70.0	0	1	0
3	61	30.0	225.0	150.0	95.0	28.58	65.0	103.0	1	1	0
4	46	23.0	285.0	130.0	84.0	23.10	85.0	85.0	1	1	0

Sous echantillonnage de la classe majoritaire unsampling

	age	cigsPerDay	totChol	sysBP	diaBP	BMI	heartRate	glucose	male	yes	yes	risque
0	39	0.0	195.0	106.0	70.0	26.97	80.0	77.0	0	0	0	0
1	46	0.0	250.0	121.0	81.0	28.73	95.0	76.0	1	0	0	0
2	48	20.0	245.0	127.5	80.0	25.34	75.0	70.0	0	1	0	0
3	61	30.0	225.0	150.0	95.0	28.58	65.0	103.0	1	1	0	1
4	46	23.0	285.0	130.0	84.0	23.10	85.0	85.0	1	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
4232	68	0.0	176.0	168.0	97.0	23.14	60.0	79.0	0	0	0	1
4233	50	1.0	313.0	179.0	92.0	25.97	66.0	86.0	0	1	0	1
4234	51	43.0	207.0	126.5	80.0	19.71	65.0	68.0	0	1	0	0
4235	48	20.0	248.0	131.0	72.0	22.00	84.0	86.0	1	1	0	0
4237	52	0.0	269.0	133.5	83.0	21.47	80.0	107.0	1	0	0	0

3301 rows × 12 columns

By Mariam Sylla

	age	cigsPerDay	totChol	sysBP	diaBP	BMI	heartRate	glucose	male	yes	yes	risque
3218	42	20.0	225.0	111.0	71.0	23.43	95.0	85.0	1	1	0	0
597	60	0.0	276.0	144.0	78.0	26.98	60.0	88.0	0	0	0	0
2692	36	20.0	177.0	115.0	63.5	22.54	71.0	73.0	1	1	0	0
2092	42	9.0	185.0	123.0	74.0	24.41	83.0	92.0	1	1	0	0
4007	46	0.0	254.0	135.0	100.0	27.86	83.0	75.0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
3521	59	0.0	190.0	127.0	77.0	28.47	80.0	100.0	0	0	0	0
2546	40	0.0	178.0	119.0	78.5	23.28	72.0	75.0	1	0	0	0
2124	62	5.0	254.0	167.5	102.5	27.15	75.0	83.0	1	1	0	0
725	40	0.0	251.0	135.0	87.0	31.60	75.0	80.0	1	0	0	0
826	50	0.0	232.0	127.5	85.0	25.09	75.0	79.0	0	0	0	0

442 rows × 12 columns

	age	cigsPerDay	totChol	sysBP	diaBP	BMI	heartRate	glucose	male	yes	yes	risque
3218	42	20.0	225.0	111.0	71.0	23.43	95.0	85.0	1	1	0	0
597	60	0.0	276.0	144.0	78.0	26.98	60.0	88.0	0	0	0	0
2692	36	20.0	177.0	115.0	63.5	22.54	71.0	73.0	1	1	0	0
2092	42	9.0	185.0	123.0	74.0	24.41	83.0	92.0	1	1	0	0
4007	46	0.0	254.0	135.0	100.0	27.86	83.0	75.0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
4221	50	0.0	260.0	119.0	74.0	21.85	80.0	72.0	0	0	0	1
4223	56	0.0	287.0	149.0	98.0	21.68	90.0	75.0	0	0	0	1
4226	58	0.0	233.0	125.5	84.0	26.05	67.0	76.0	0	0	0	1
4232	68	0.0	176.0	168.0	97.0	23.14	60.0	79.0	0	0	0	1
4233	50	1.0	313.0	179.0	92.0	25.97	66.0	86.0	0	1	0	1

884 rows × 12 columns

```
risque
0    0.5
1    0.5
Name: proportion, dtype: float64
```

Les deux classes sont bien équilibrées

Séparation en base de train et test (Data and Target Split)

By Mariam Sylla

```
      age  cigsPerDay  totChol  sysBP  diaBP    BMI  heartRate  glucose  male  \
571    50         0.0    230.0  133.0   91.0  25.74       72.0     70.0     0
1248   49         9.0    266.0  159.0   88.0  20.66       76.0     84.0     1
310    47         0.0    250.0  114.0   77.0  24.16       80.0     93.0     1
1553   67        15.0    285.0  155.0   90.0  30.42       70.0     77.0     0
1888   63        30.0    225.0  146.0   82.0  27.17       70.0     85.0     0
...   ...         ...      ...    ...    ...    ...        ...      ...   ...
1095   51        20.0    219.0  125.0   71.0  21.19       77.0     75.0     0
3835   62         0.0    266.0  124.0   69.0  22.90       66.0     82.0     1
1852   50         0.0    229.0  105.0   72.5  26.25       90.0     79.0     1
663    59        20.0    206.0  167.0   89.5  25.83       72.0     75.0     1
4111   61         0.0    257.0  141.0   80.0  33.90       85.0     60.0     1

      yes  yes
571     0    0
1248    1    0
310     0    0
1553    1    0
1888    1    0
...   ...  ...
1095    1    0
3835    0    0
1852    0    0
663     1    0
4111    0    0

[618 rows x 11 columns]
```

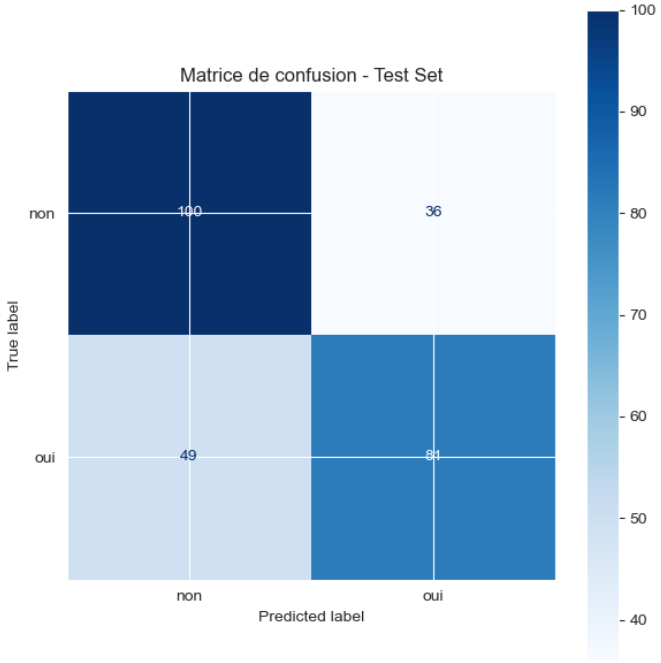
LogisticRegression

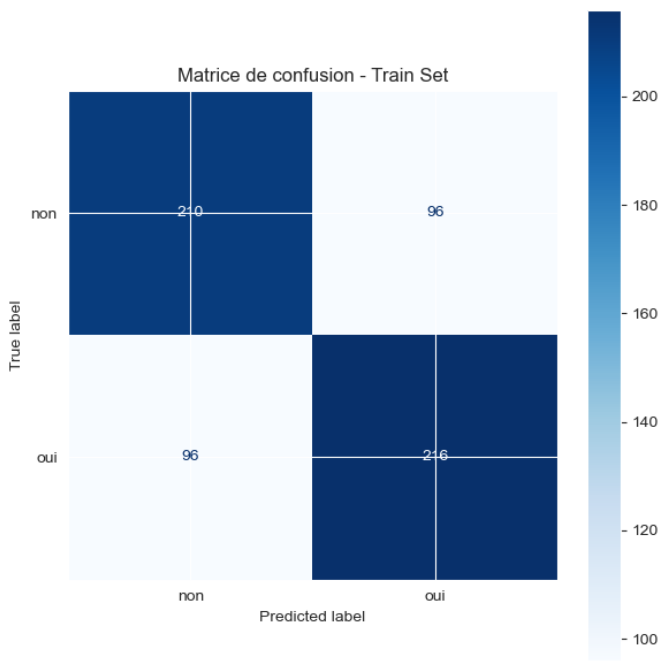
```
LogisticRegression(solver='liblinear')
```

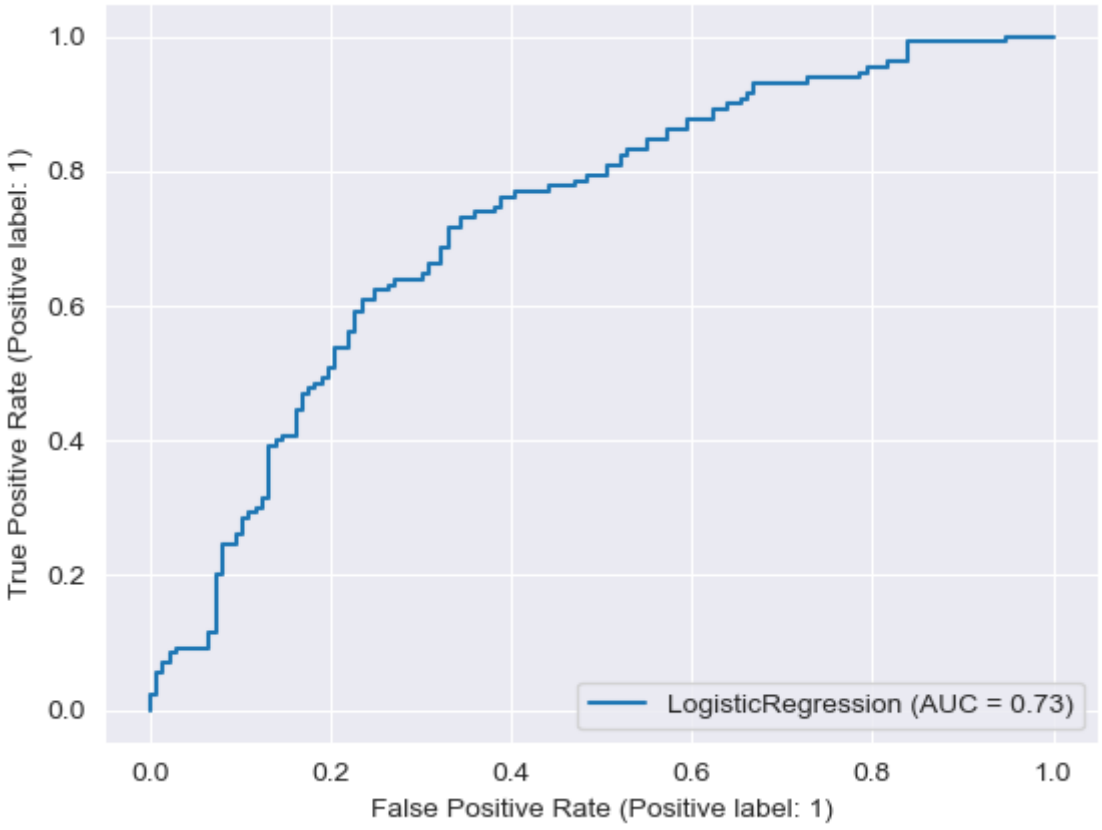
```
Performance sur la base d'entrainement
accuracy score train: 0.6893203883495146
Train Classification report :
              precision    recall  f1-score   support

           0       0.69      0.69      0.69       306
           1       0.69      0.69      0.69       312

    accuracy                           0.69       618
   macro avg       0.69      0.69      0.69       618
weighted avg       0.69      0.69      0.69       618

Performance sur la base de test
accuracy score test: 0.6804511278195489
Test Classification report :
              precision    recall  f1-score   support

           0       0.67      0.74      0.70       136
           1       0.69      0.62      0.66       130

    accuracy                           0.68       266
   macro avg       0.68      0.68      0.68       266
weighted avg       0.68      0.68      0.68       266
```

Sur la base test Nous avons un accuracy de 68%, le modele a une performance de 68%,ce qui veut le modèle fait une bonne prediction dans 68% des cas.

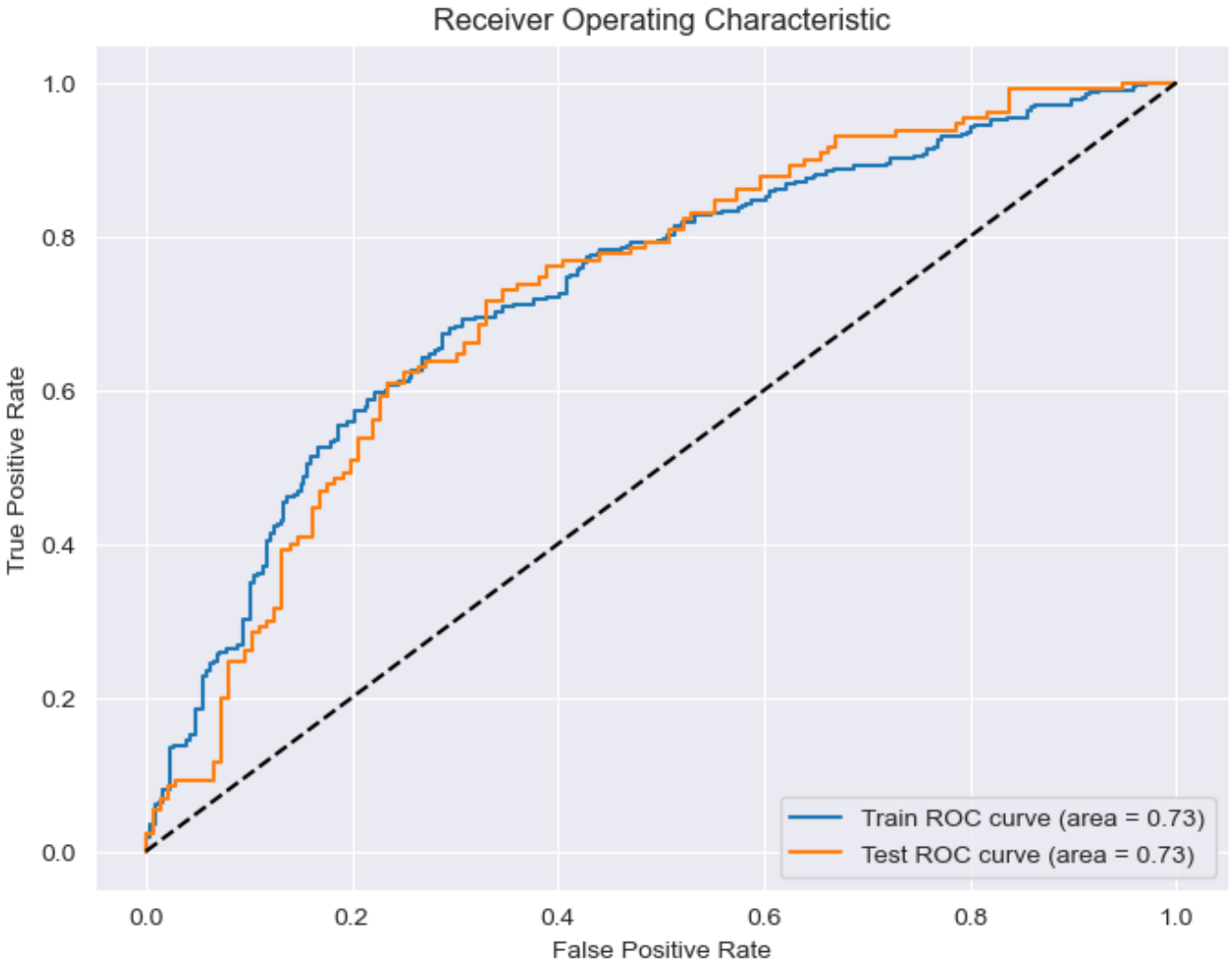
By Mariam Sylla







By Mariam Sylla



Nous avons un AUC de 73% ce qui veut dire que le modèle est capable de prédire dans 73% des cas si le patient a le risque de maladie cardiaque

Utilisation du modele

By Mariam Sylla

```
Entrez l'age du patient:
>>>42
Combien de cigarette le patient fume par jour?
>>>20
Quel est le taux de cholesterol du patient?
>>>225
Quelle est la tension arterielle du patient?
>>>111
Quelle est la pression arterielle du patient?
>>>71
Quel est l'indice de la masse corporelle du patient?
>>>24
Quelle est la frequence cardiaque du patient?
>>>95
Quelle est le taux de glucose du patient?
>>>85
Le patient est-il un homme?
>>>1
Est ce que le patient fume actuellement?
>>>1
Le patient a-t-il le diabete?
>>>0
Le patient de 42 ans n'a pas de risque de maladie cardiaque
```

By Mariam Sylla