ANALISI DI UN DATASET
RELATIVO A COSTI ASSICURATIVI
PER RISCHI SANITARI

Corso di Data Analyst IFOA - Epicode

Naira Koehler Cammarata

Roma, 29/01/2024

INDICE

1. Obiettivi

Il presente report si propone di esplorare approfonditamente il dataset riguardante costi assicurativi sostenuti dalla popolazione statunitense in relazione a rischi sanitari[1], fornendo un'analisi dettagliata delle variabili ivi presenti.

Il dataset in esame riporta, per ciascun soggetto della popolazione assicurata (il campione ricomprende 1.338 persone), i seguenti dati: **(i)** l'età; **(ii)** il sesso; **(iii)** l'indice di massa corporea (BMI); **(iv)** il numero di figli; **(v)** la qualità o meno di fumatore; **(vi)** l'area geografica di residenza.

A esito dell'attività di analisi, saranno espresse talune valutazioni in merito alle correlazioni tra le variabili in esame e i costi assicurativi nel settore sanitario, evidenziando quali sono i fattori maggiormente incisivi, con l'obiettivo di consentire l'assunzione di decisioni consapevoli da parte degli operatori economici.

2. Analisi dei dati

In via preliminare, si riportano nella seguente tabella, per ciascuna delle tipologie di dati presi in considerazione, la media, la deviazione standard, il valore minimo, i quartili (25%, 50% e 75%) e il valore massimo.

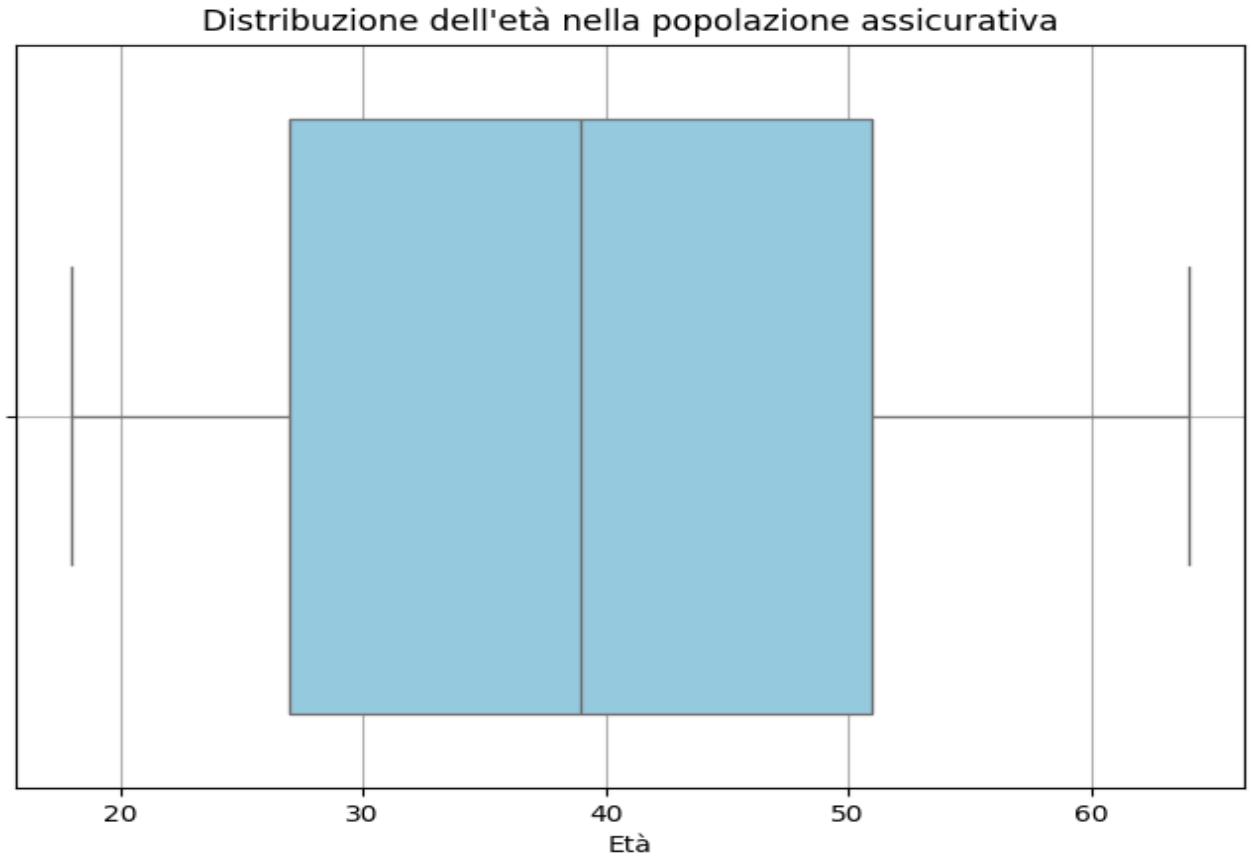
	Età	**Sesso**	**BMI**	**Figli**	**Qualità di fumatore**	**Costo Assicurazione**
Media	39,21	0,51	30,66	1,09	0,20	13.270,42
Deviazione Standard	14,05	0,50	6,10	1,21	0,40	12.110,01
Valore minimo	18,00	0,00	15,96	0,00	0,00	1.121,87
Pimo quartile (25%)	27,00	0,00	26,30	0,00	0,00	4.740,29
Secondo quartile (50%)	39,00	1,00	30,40	1,00	0,00	9.382,03
Terzo quartile 75%	51,00	1,00	34,69	2,00	0,00	16.639,91
Valore massimo	64,00	1,00	53,13	5,00	1,00	63.770,43

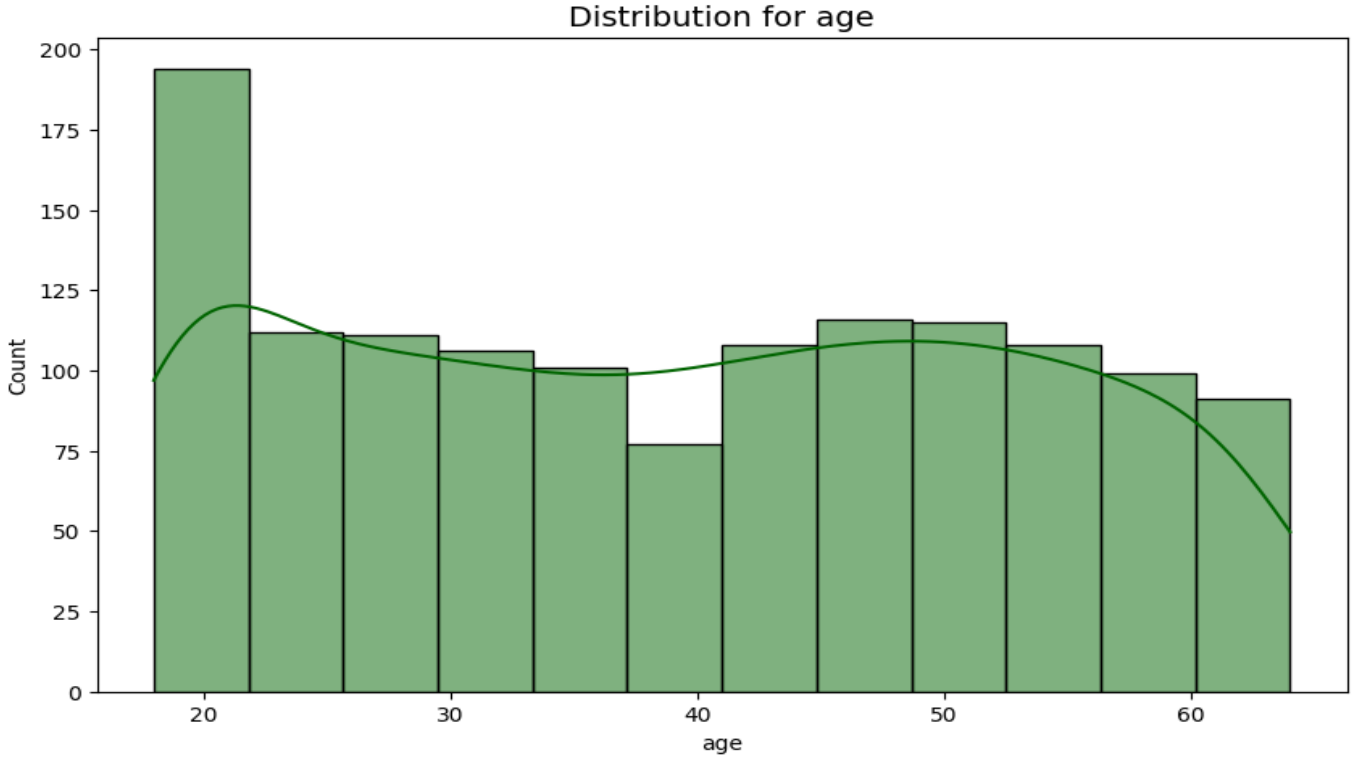
2.1. Età

L'età media della popolazione assicurativa è di circa 39 anni, con un'età minima di 18 anni e un'età massima di 64 anni. Possiamo notare una maggioranza di individui di mezza età, con una deviazione standard di circa 14 anni.

Di seguito, si riportano rappresentazioni grafiche della distribuzione dell'età nella popolazione assicurata sotto forma di boxplot e istogramma.

[1] Il relativo file, denominato "insurance.csv", è stato scaricato dal sito https://www.kaggle.com/datasets/ahmettezcantekin/beginner-datasets?resource=download.

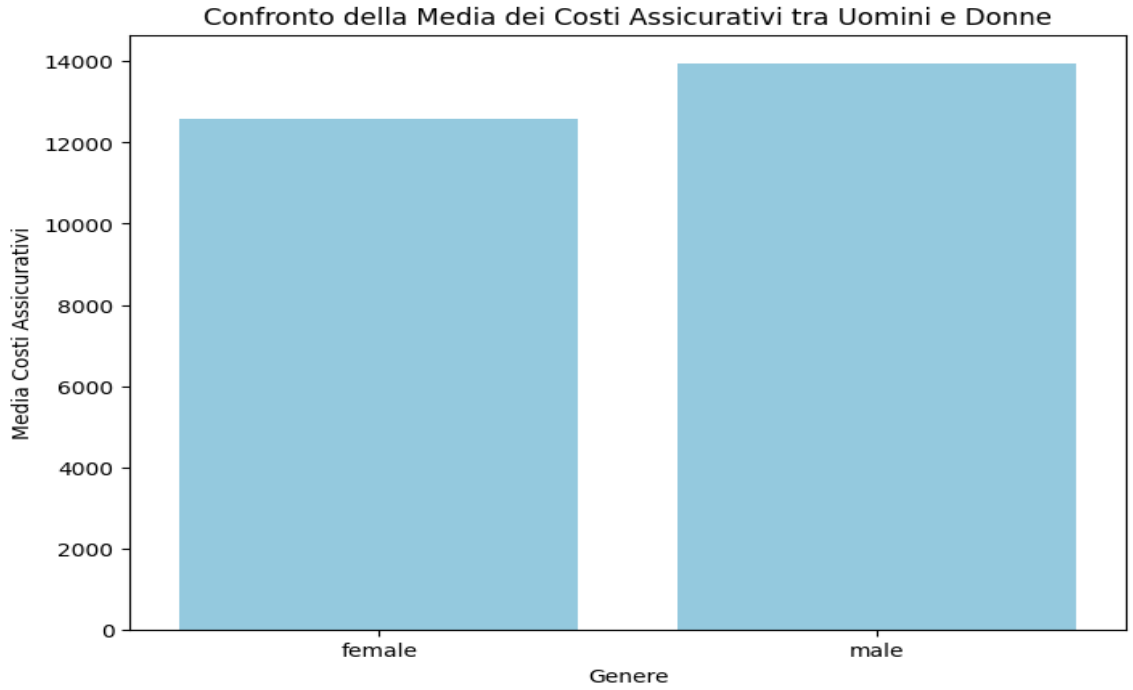

Distribuzione dell'età nella popolazione assicurativa
20
30
40
50
60
Età

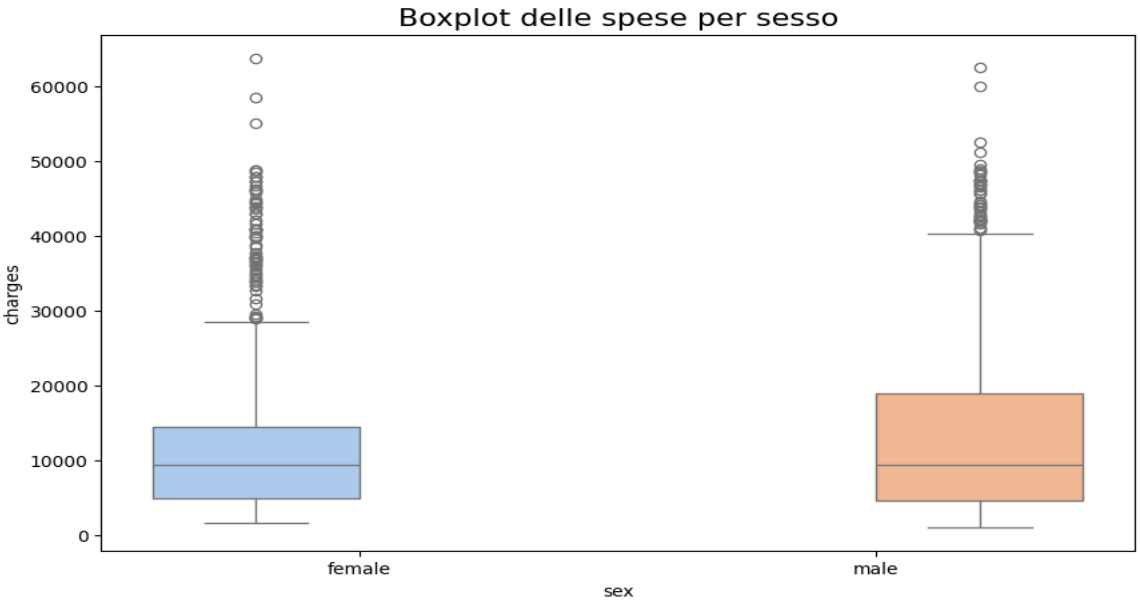

Distribution for age
200
175
150
125
100
75
50
25
0
Count
20
30
40
50
60
age

2.2. Sesso

Il 50,5% dei partecipanti sono uomini. La codifica utilizzata è 0 per le donne e 1 per gli uomini. L'analisi della distribuzione dei costi assicurativi in base al sesso rivela la seguente differenza: la media dei costi per le donne ammonta a $12.569,58, mentre la media dei costi per gli uomini a $13.956,75, per una differenza pari a $ 1.387,17.

Di seguito, si riportano rappresentazioni grafiche della distribuzione dei costi assicurativi tra uomini e donne. Al riguardo, si sottolinea che i boxplot sotto riportato evidenzia come la mediana dei costi per sesso non si discosti molto dalla media, con una distribuzione abbastanza equa per le donne a differenza degli uomini, dove si nota che gli stessi sono in maggioranza a partire dalla mediana (50%) fino al 3° percentile (75%). Inoltre, si registra una grande quantità di outliers nei due boxplot.



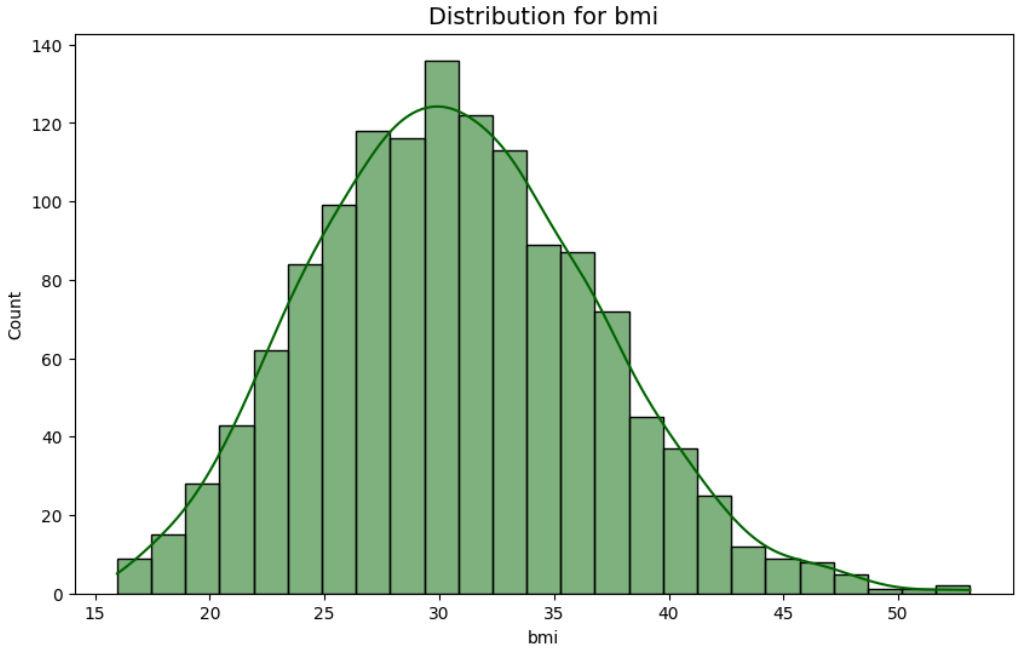


2.3. BMI (Indice di Massa Corporea):

Il BMI medio nella popolazione assicurativa è di circa 30,66, con una deviazione standard di circa 6,10. Si evidenzia una distribuzione normale (gaussiana), ma avendo valore 30 possiamo dire che la popolazione analizzata è in sovrappeso.

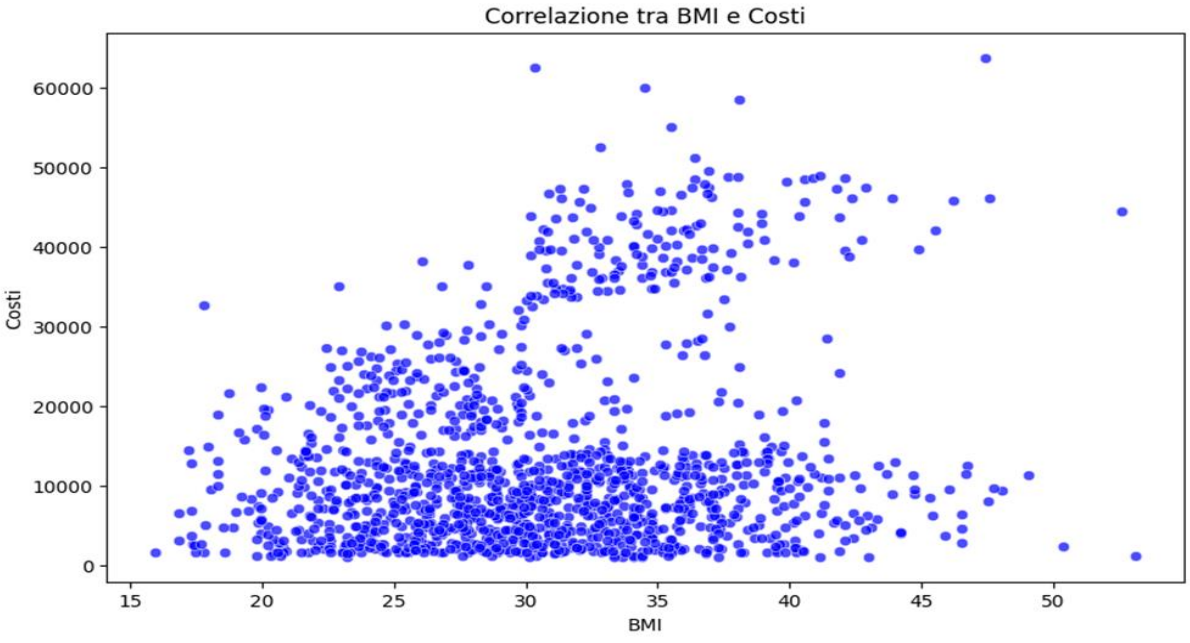
La distribuzione del BMI è eterogenea, con valori compresi tra 15,96 e 53,13.

Di seguito, si riporta una rappresentazione grafica della distibuzione della popolazione assicurata in relazione ai valori di BMI.



L'indice di correlazione tra BMI e costi assicurativi è di 0,2. La correlazione è leggera, indicando una relazione crescente moderata.

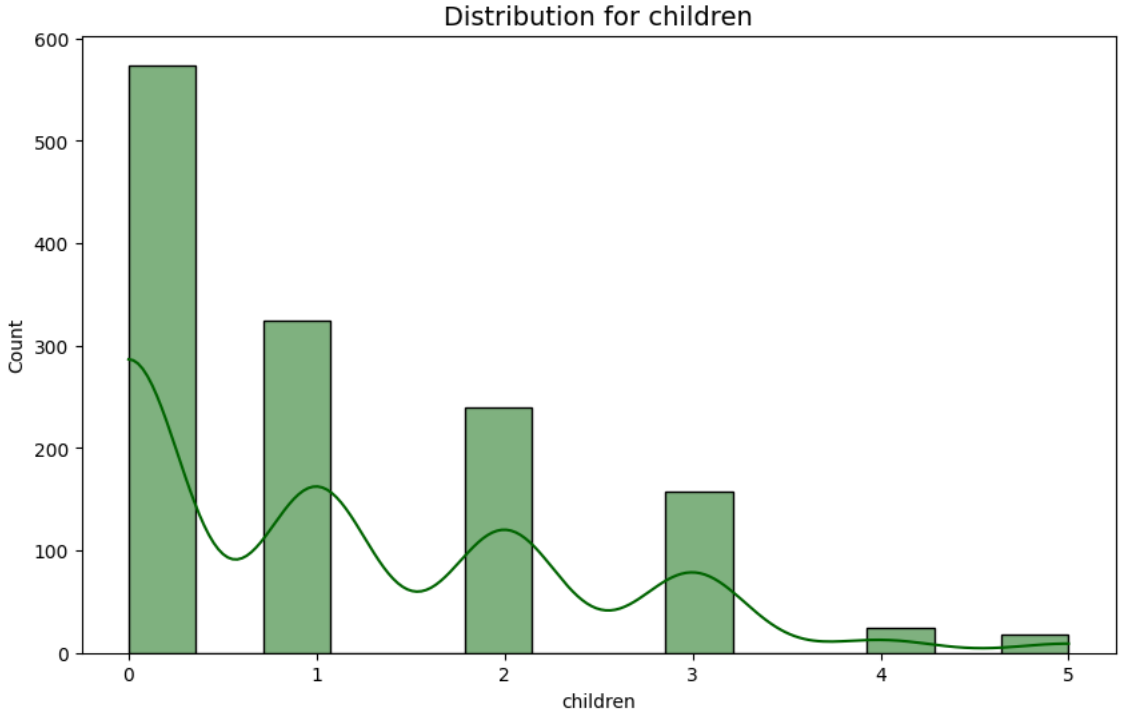
Si riporta di seguito un grafico di dispersione che riporta la citata correlazione.



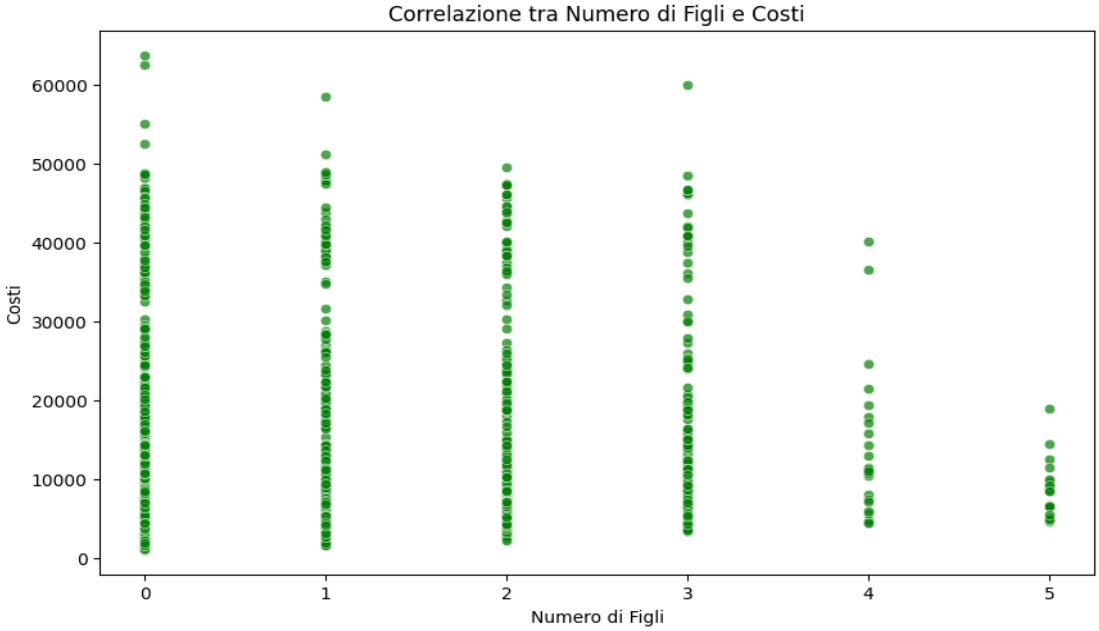
2.4. Figli

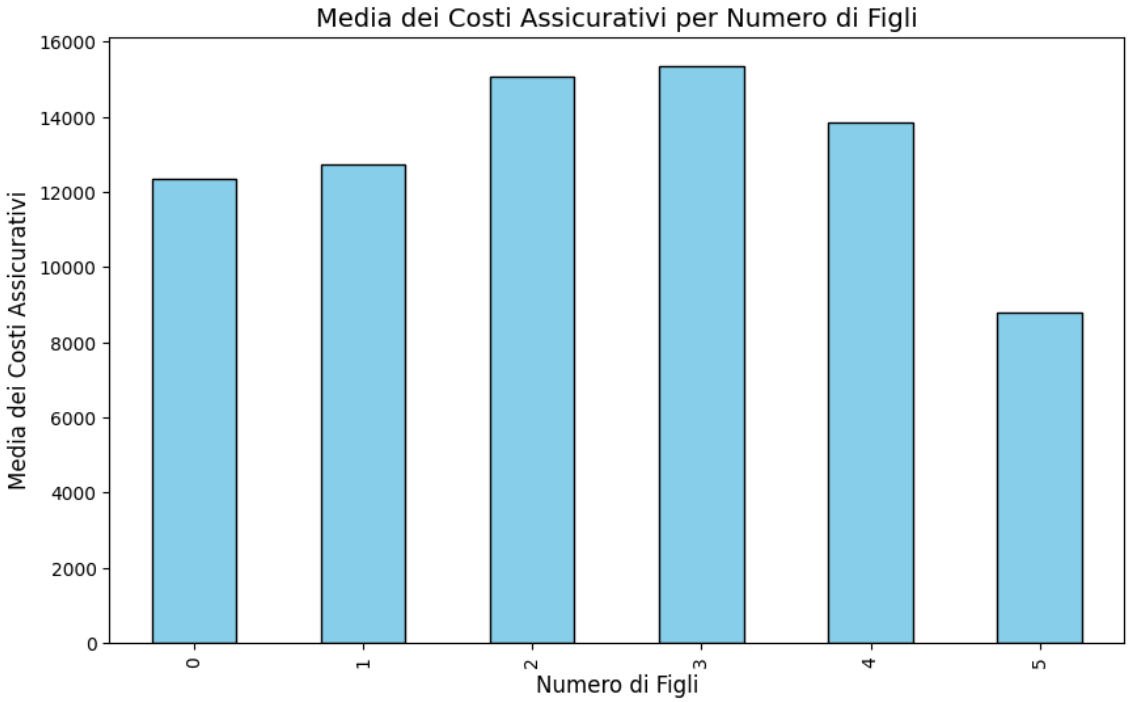
La media di figli per partecipante è di circa 1,09, con un massimo di 5 figli.

Come si vede dall'istogramma sottoriportato, che mette in relazione la popolazione assicurata con il numero di figli, la maggioranza della popolazione ne è priva.



L'indice di correlazione tra costi assicurativi e figli, di cui all'istogramma che segue, tra costi e figli è pari a 0,07 e, pertanto, è trascurabile.



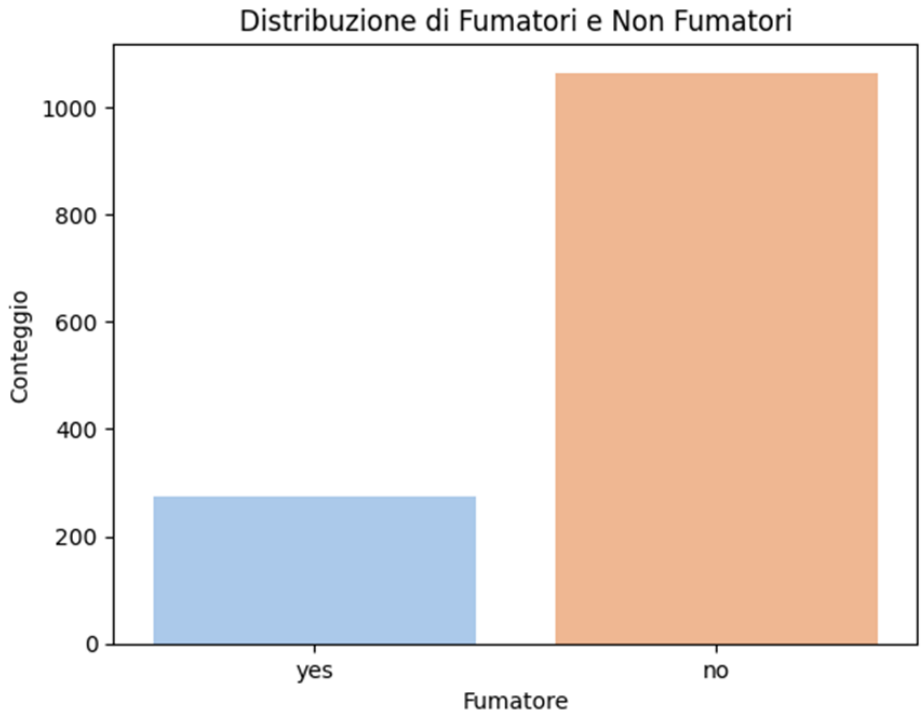


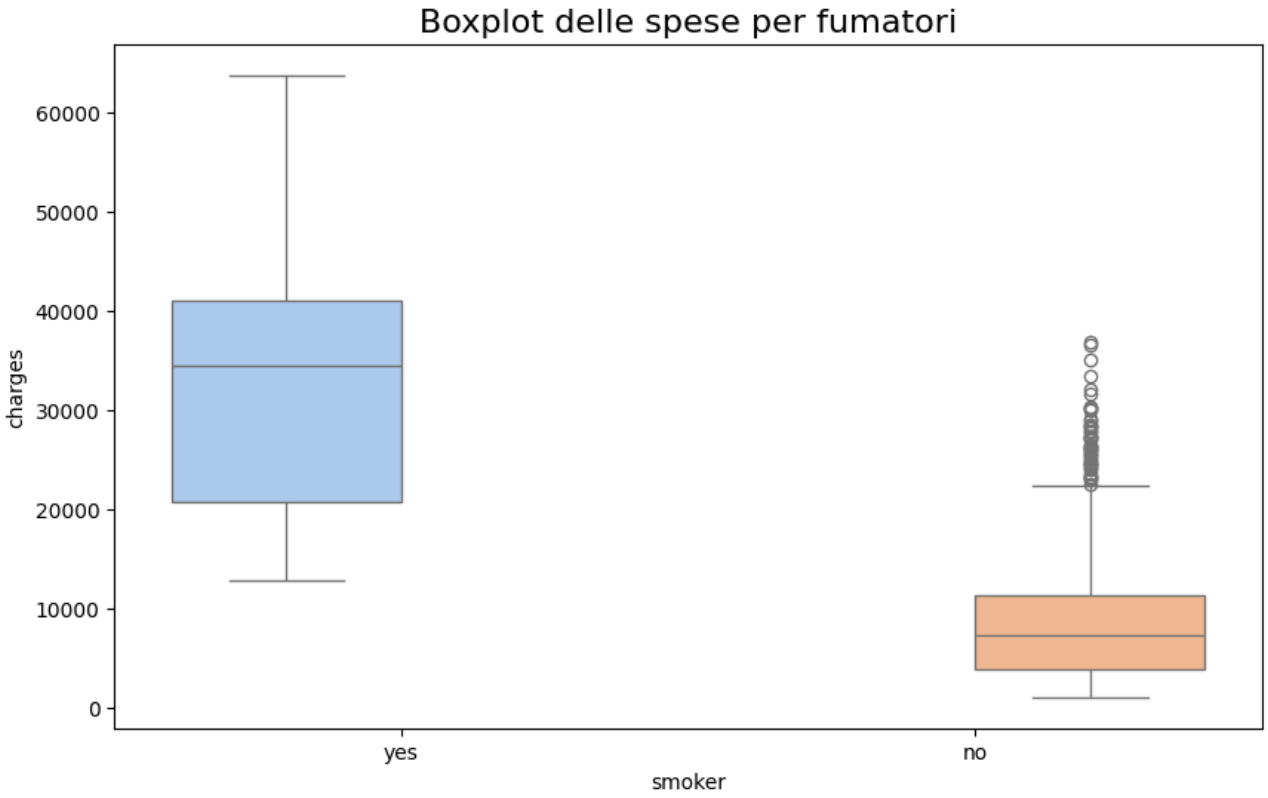
2.5. Qualità di fumatore

Circa il 20,5% della popolazione è costituito da fumatori. La codifica utilizzata è 0 per i non fumatori e 1 per i fumatori.

L'indice di correlazione tra costi assicurativi e qualità di fumatore è parri a 0,8. La correlazione è, quindi, significativa, indicando una relazione crescente forte.

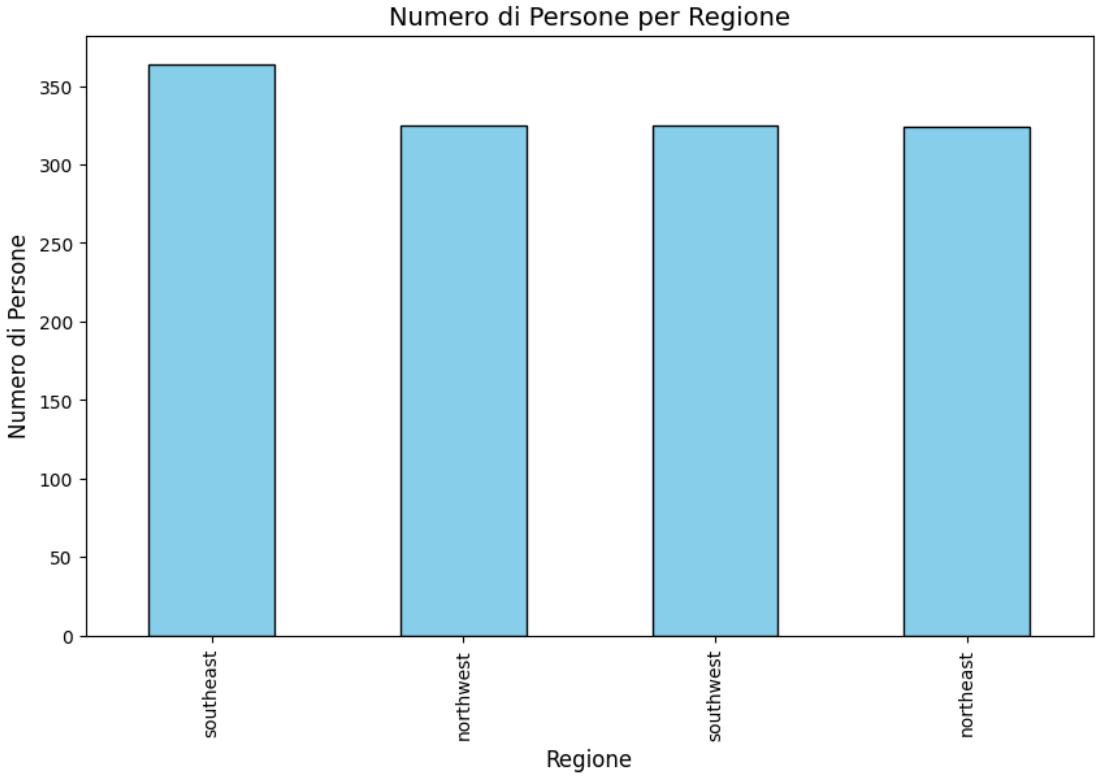
In particolare, la media dei costi assicuativi per non fumatori è pari a $ 8.434,27, mentre la media dei costi per non fumatori è pari a $ 32.050,23, con una differenza di ben $ 23.615,96.





2.6. Aree geografiche di provenienza

Come si può osservare dall'istogramma che segue, la popolazione oggetto di analisi è equamente distribuita nelle quattro aree geografiche di riferimento (['southwest', 'southeast', 'northwest', 'northeast']). Ne consegue la pressoché trascurabile correlazione tra tale dato e i costi assicurativi.

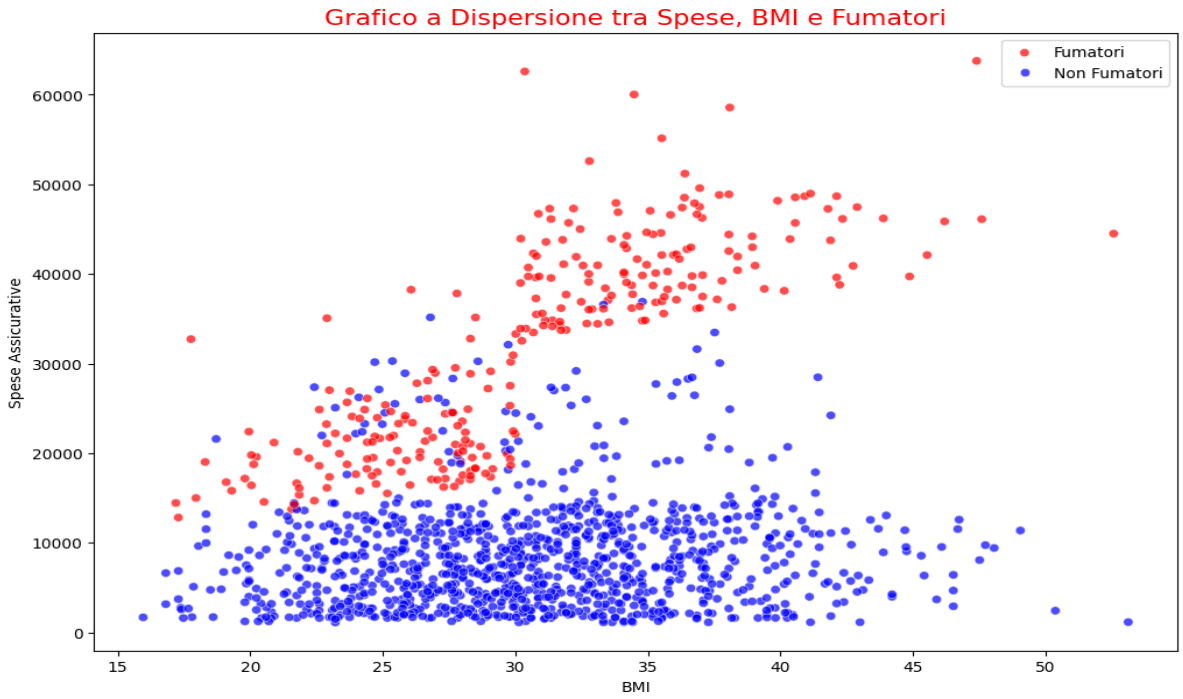


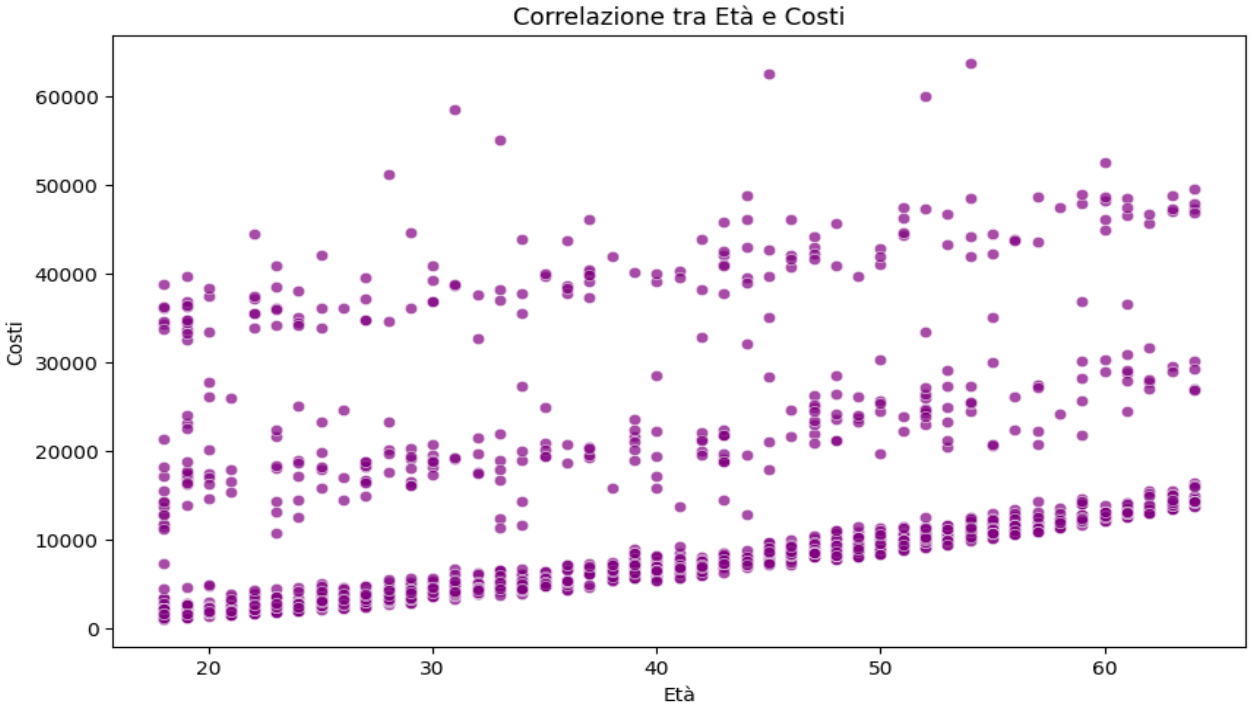
2.7. Costi Assicurativi

La media dei costi assicurativi è di circa $ 13.270,42, con una deviazione standard di circa $ 12.110,01. I costi variano da un minimo di $ 1.121,87 a un massimo di $ 63.770,43.

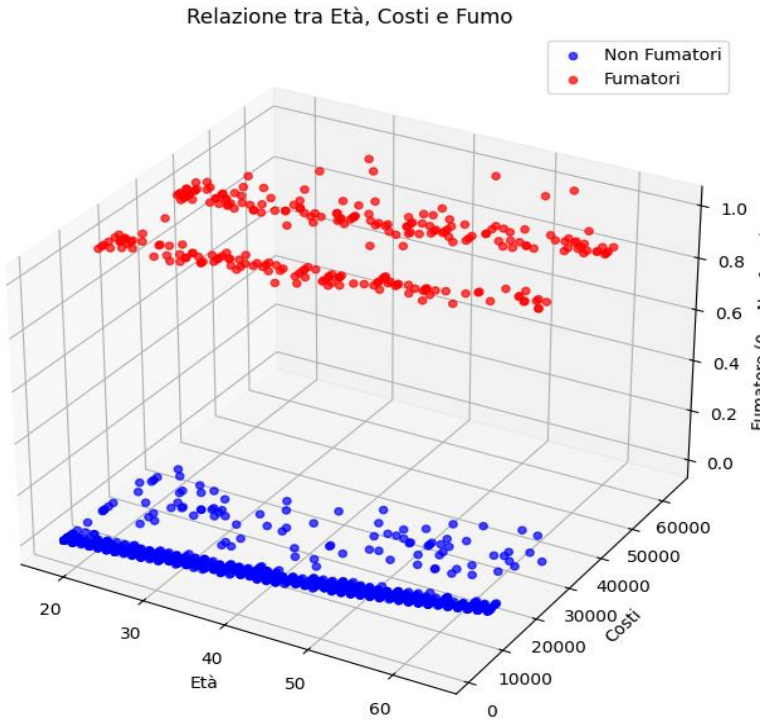
Di seguito, si riportano due grafici di dispersione, rispettivamente:

- tra spese, BMI e qualità di fumatori;
- tra età e costi assicurativi.



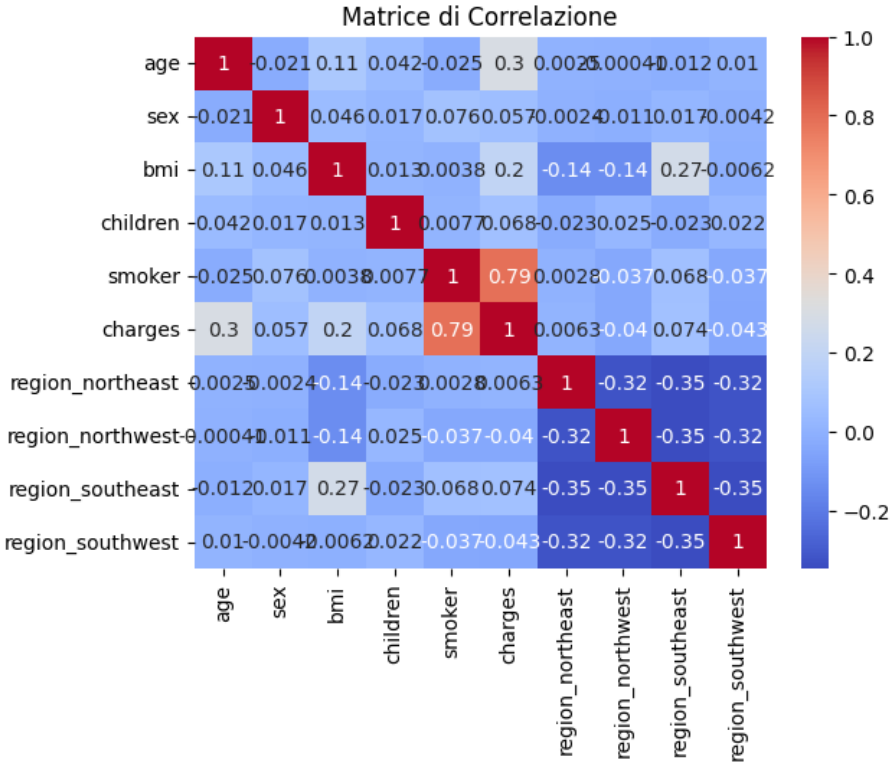


Nel seguente grafico a dispersione in 3D, è possibile osservare la correlazione tra costi, età e qualità di fumatori.

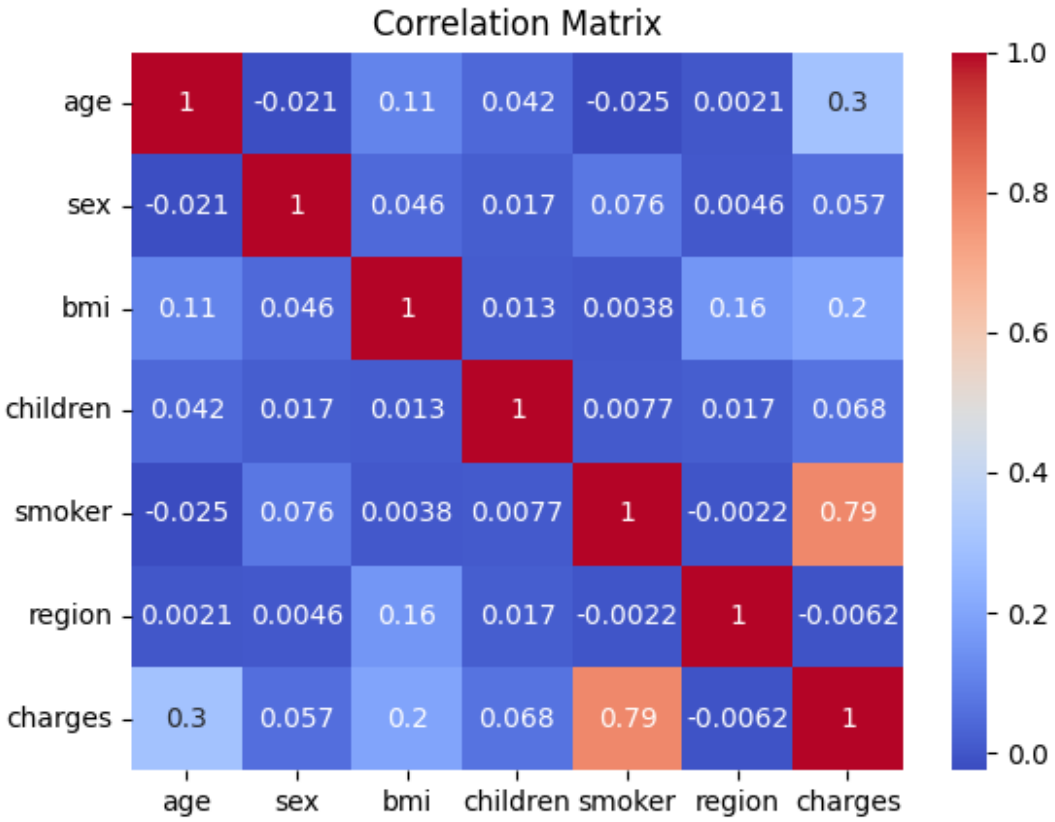


3. Conclusioni

Al fine di meglio illustrare le conclusioni del presente report è stata predisposta una matrice di correlazione tra i dati oggetto di analisi (1: max correlazione; 0: nessuna correlazione; -1: max correlazione negativa).



Per comodità di consultazione, si riporta di seguito la matrice di correlazione, senza indicazione delle aree geografiche di provenienza (che rappresentano il fattore, senz'altro, meno rilevante).



Alla luce di quanto precede, è possibile concludere, in estema sintesi, che:

- i fattori che incidono maggiormente sui costi assicurativi consistono, nell'ordine, nella qualità di fumatore (corr.: 0,79), nell'età (corr.: 0,3) e nel BMI (corr.: 0,2);
- i fattori che, invece, si possono ritenere trascurabili sono dati dal possesso di figli (corr.: 0,068), dal sesso (corr.: 0,057), nonché dall'area geografica di provenienza (corr.: -0,0062).