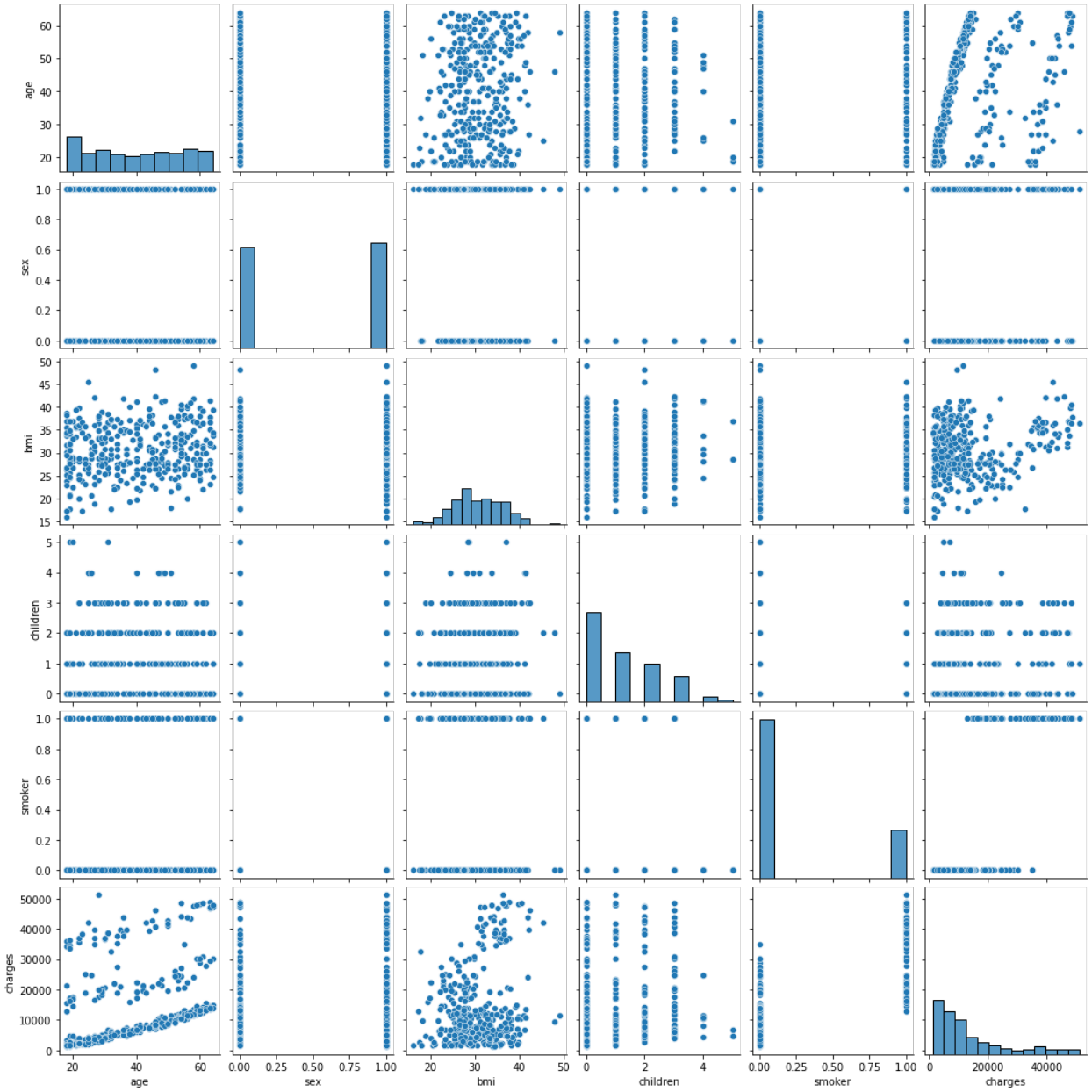
Andy Castiilo 18040
Mineria de Datos

Laboratorio 4

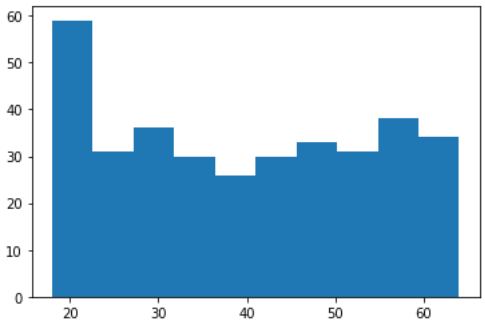
Información del dataset:
El dataset tienen un total de 7 columnas (age, sex, bmi, children, smoker, región, charges) con un total de 348 filas. Todas las columnas son de tipo numérico 2 float y el resto int.

Hipótesis y objetivo:
Lo que se quiere realizar en este laboratorio es poder predecir, en este caso charges, en base al bmi y usando regresión lineal,

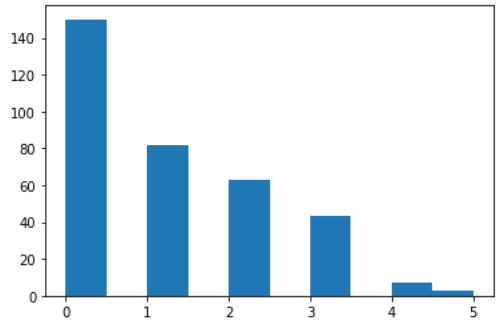
Solución y exploración:
Se analizó la relación de todos los datos a través de histogramas y pairplots.



Pairplot de los datos



Histograma del campo de edades



Histograma del campo de children

Luego se observó si había datos que fueron nulos para poder ver si hay que hacer algo con estos tipos de datos:

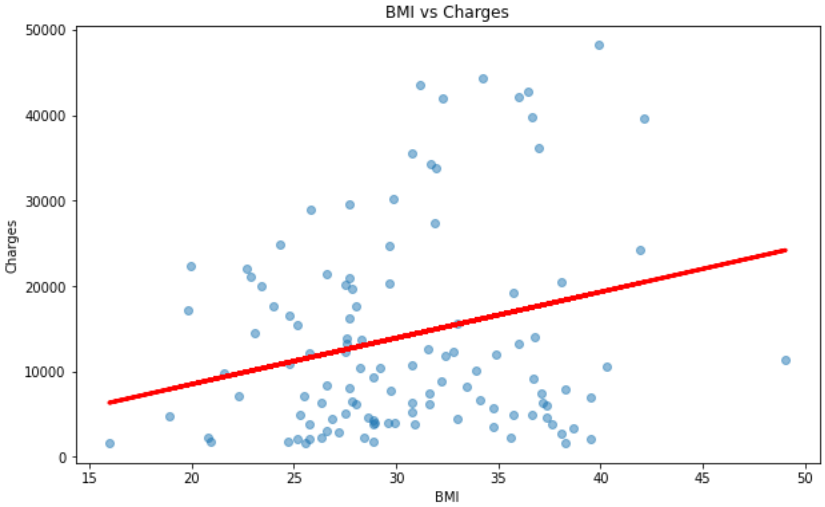
Column	Non-Null Count		
age	348 non-null	age	348
sex	348 non-null	sex	348
bmi	348 non-null	bmi	348
children	348 non-null	children	348
smoker	348 non-null	smoker	348
region	348 non-null	region	348
charges	348 non-null	charges	348

Como se puede ver en las tablas anteriores hay un total de 348 datos por cada columna y todos son non-null por lo que no nos tenemos que preocupar por los datos nulos.

Por último se dividió la data en dos sets diferentes, una de training para poder entrenar al modelo y otro set de test para poder ver si el entrenamiento funcionó.

Resultados:

Luego de haber entrenado el modelo, se procedió a graficar la regresión lineal.



Luego se sacó el valor de R^2 y dio un total de -11 por lo que se puede decir que al utilizar estos datos y entrenarlos no dio un buen resultado con la región lineal e intentar predecir charges en base al bmi.