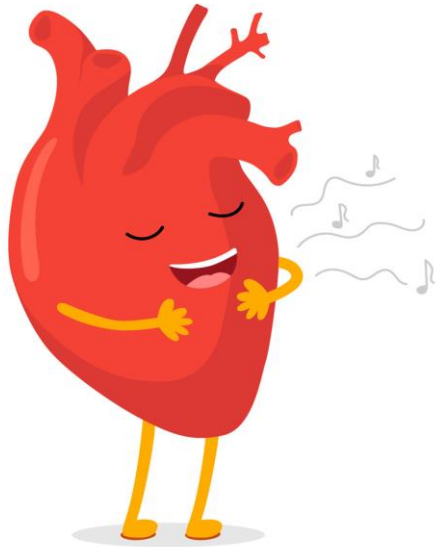


CODERHOUSE

Heart Attack Risk Prediction

¿Qué factores influyen en el Riesgo Cardíaco?

Martinez, Mario Alejandro

INDICE:

CONTEXTO Y **OBJETIVOS**

Contexto

Los infartos, y en general, las enfermedades cardiovasculares son la principal causa de muerte tanto en hombres como en mujeres en todo el mundo. Una detección temprana de esta cardiopatía aumentaría las posibilidades de tratamiento y de prevención. Estas patologías no se producen por una única causalidad sino que *existen muchos factores de riesgo dentro de los cuales podríamos mencionar otras patologías*, como la hipertensión arterial o la diabetes, *antecedentes familiares o propios respecto a cardiopatías y hábitos* relacionados al ejercicio, el descanso y al día a día de los pacientes.

Un ataque cardíaco se define como la necrosis isquémica del corazón, generalmente causada por una obstrucción de las arterias que lo irrigan.

Objetivo

El objetivo general del proyecto será entonces poder desarrollar un modelo desarrollar un modelo de clasificación para predecir un ataque cardíaco.

PREGUNTAS DE **INTERÉS**

Preguntas principales o primarias

- ¿Existen uno o más factores de riesgo en los pacientes que su presencia está relacionada al riesgo de ataque cardíaco?

Preguntas secundarias (nos ayudaran a contestar las principales)

- ¿Cuál es la distribución de nuestra muestra según el riesgo cardíaco?
- ¿Puede considerarse el género como una factor de riesgo para el ataque cardíaco?
- ¿Tiene alguna relación la ubicación geográfica de los pacientes con su patología?
- ¿Cómo se distribuye en cada continente el riesgo de ataque cardíaco?
- ¿Qué factores tienen una implicancia directa con el riesgo cardíaco?
- ¿Existe alguna relación demostrable entre los factores observados en la muestra?

ANÁLISIS
EXPLORATORIO

¿Cómo es nuestro ‘**Dataset’?**

- Posee un total de 8763 registros.

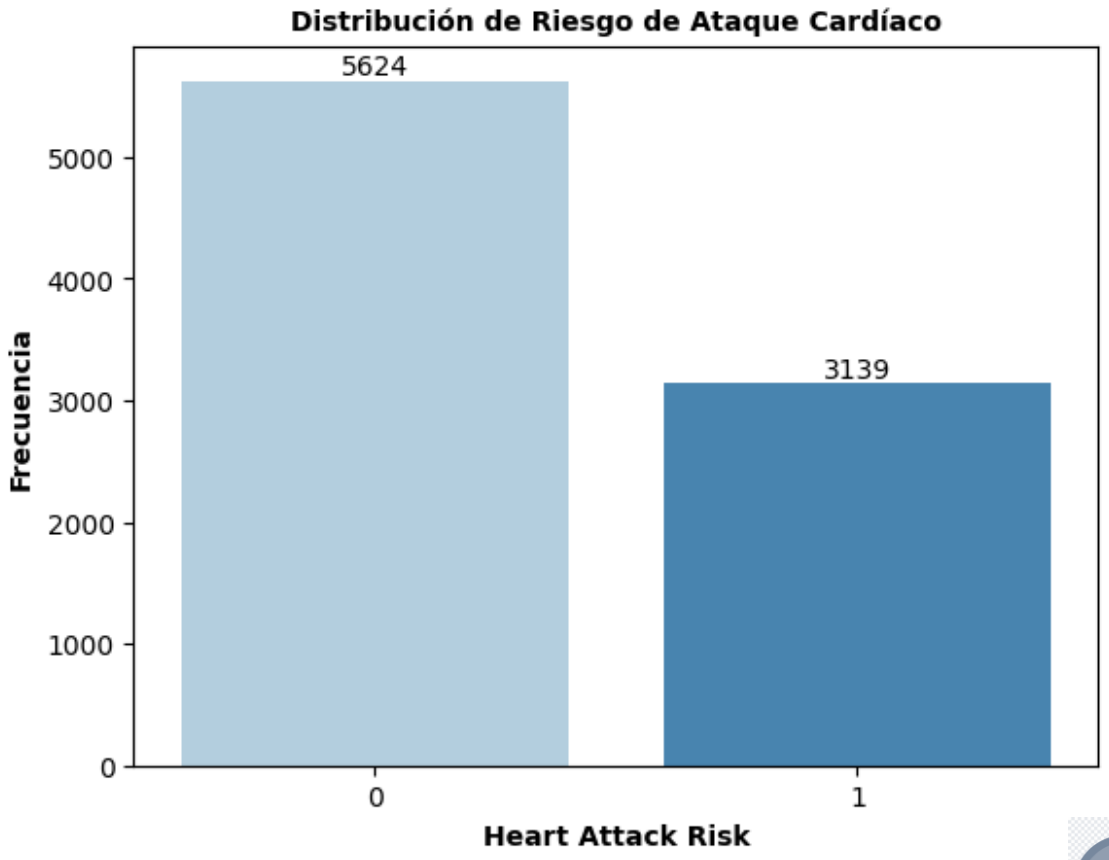
Las variables se dividen en un total de 26 columnas de las cuáles hemos ido trabajando y concluyendo que:

- Presenta una variable índice;,
- Presenta 12 variables númericas, cuantitativas tanto enteras como decimales;
- Presenta 13 variables categóricas tanto binarias como ordinales y nominales.

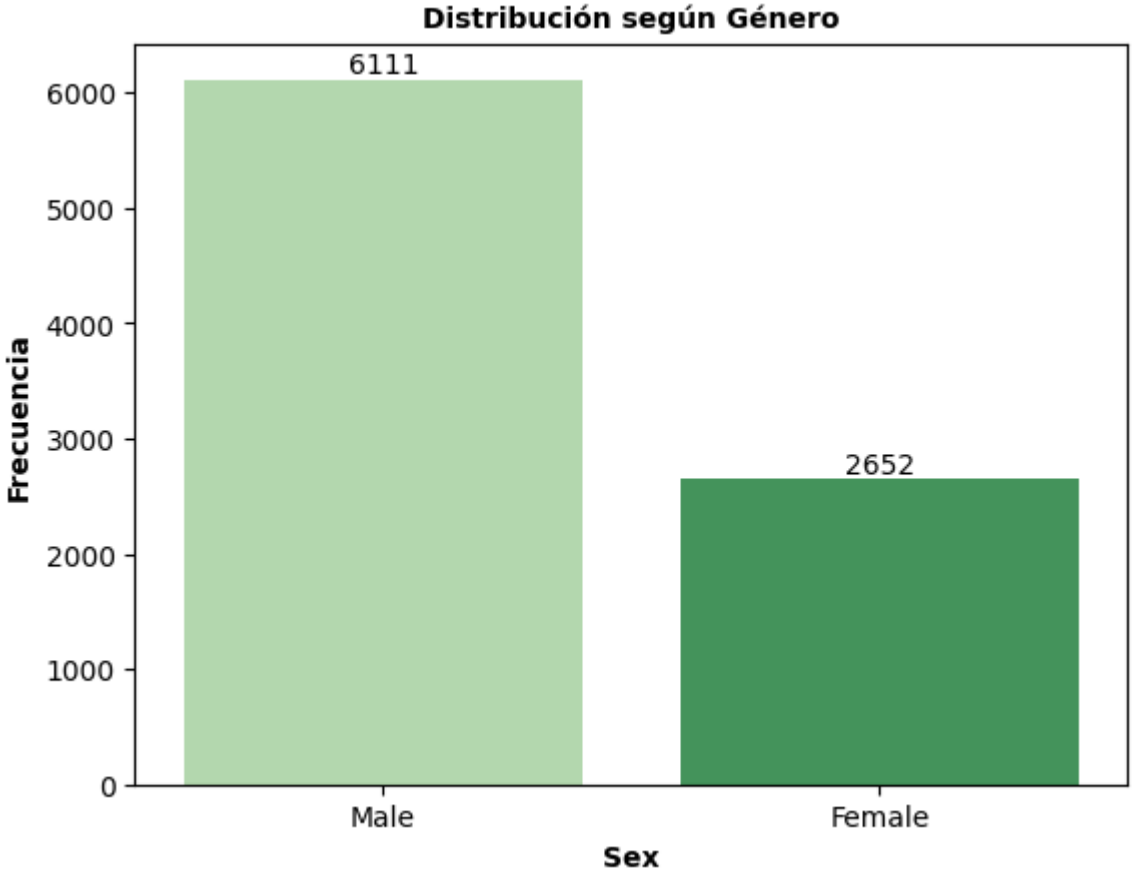
Les invitamos a conocer nuestro **Diccionario de datos para un mejor entendimiento de nuestro dataset.**

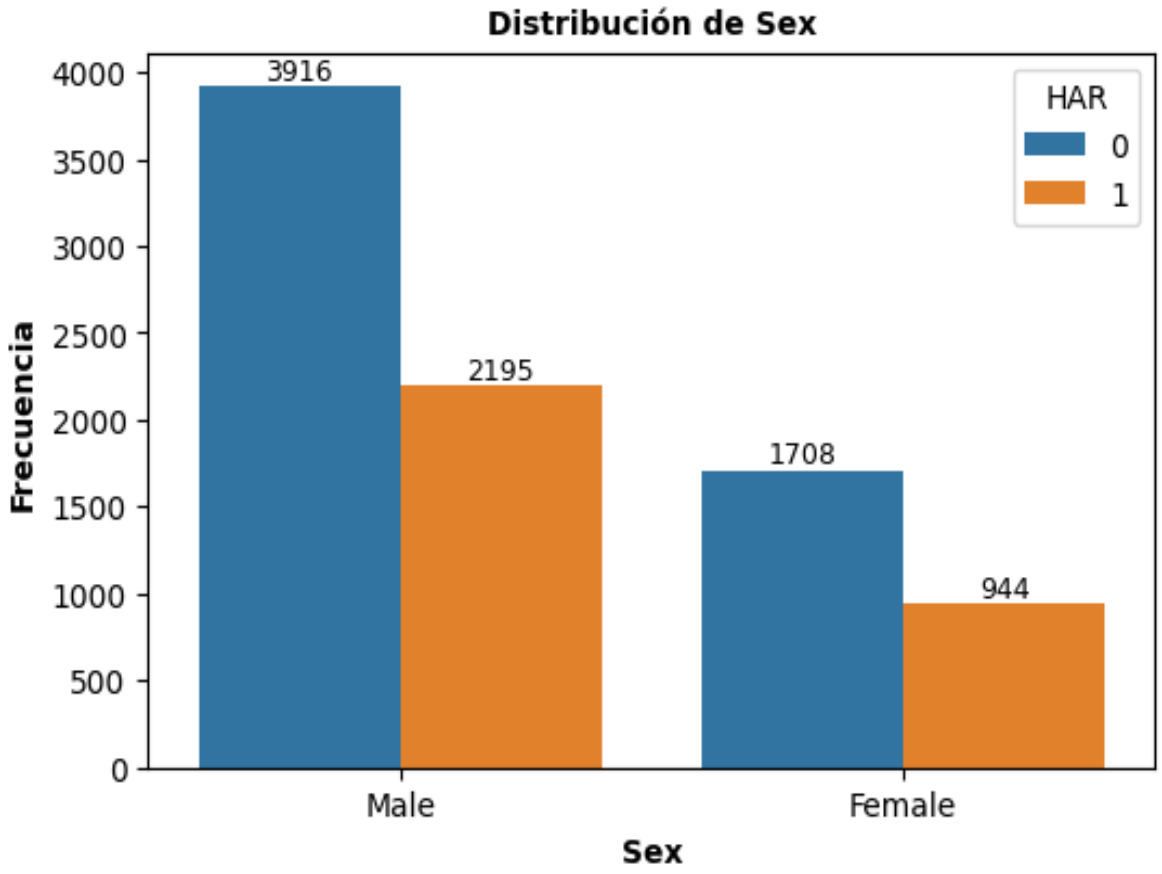
A continuación les mostraremos las frecuencias obtenidas para cada valor único presentado de la variable **‘Heart Attack Prediction’, variable target** que indica la condición médica de los pacientes en la muestra:

Clasificación de nuestra variable ‘Target’



¿Cuál es la relación entre el **Género y el Riesgo de Ataque Cardíaco?**

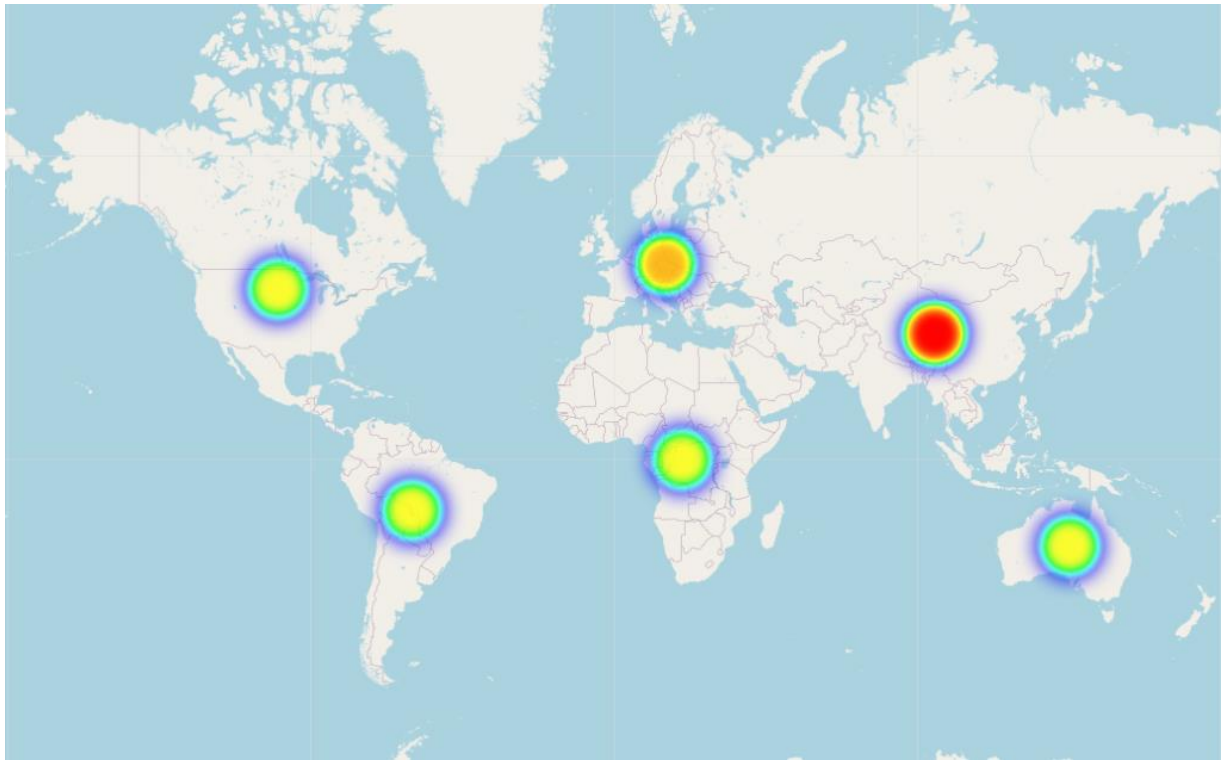




Mientras que el sexo Masculino predomina representando el 69,7% de la muestra, en ambos casos el Riesgo de Ataque Cardíaco aproximadamente el 36% (Ver sección Variables Categóricas en relación a variable target para más información)

¿Cuál es la relación entre el **Continente y el Riesgo de Ataque Cardíaco?**

Value	Count	Frequency (%)
Asia	2543	29.0%
Europe	2241	25.6%
South America	1362	15.5%
Oceania	884	10.1%
Africa	873	10.0%
North America	860	9.8%



Si bien en un primer análisis observamos que los continentes de Asia y Europa tienen una mayor presencia en nuestro dataset, al analizar el impacto junto a la variable target nos encontramos que…

¿Cuál es la relación entre el **Continente y el Riesgo de Ataque Cardíaco?**

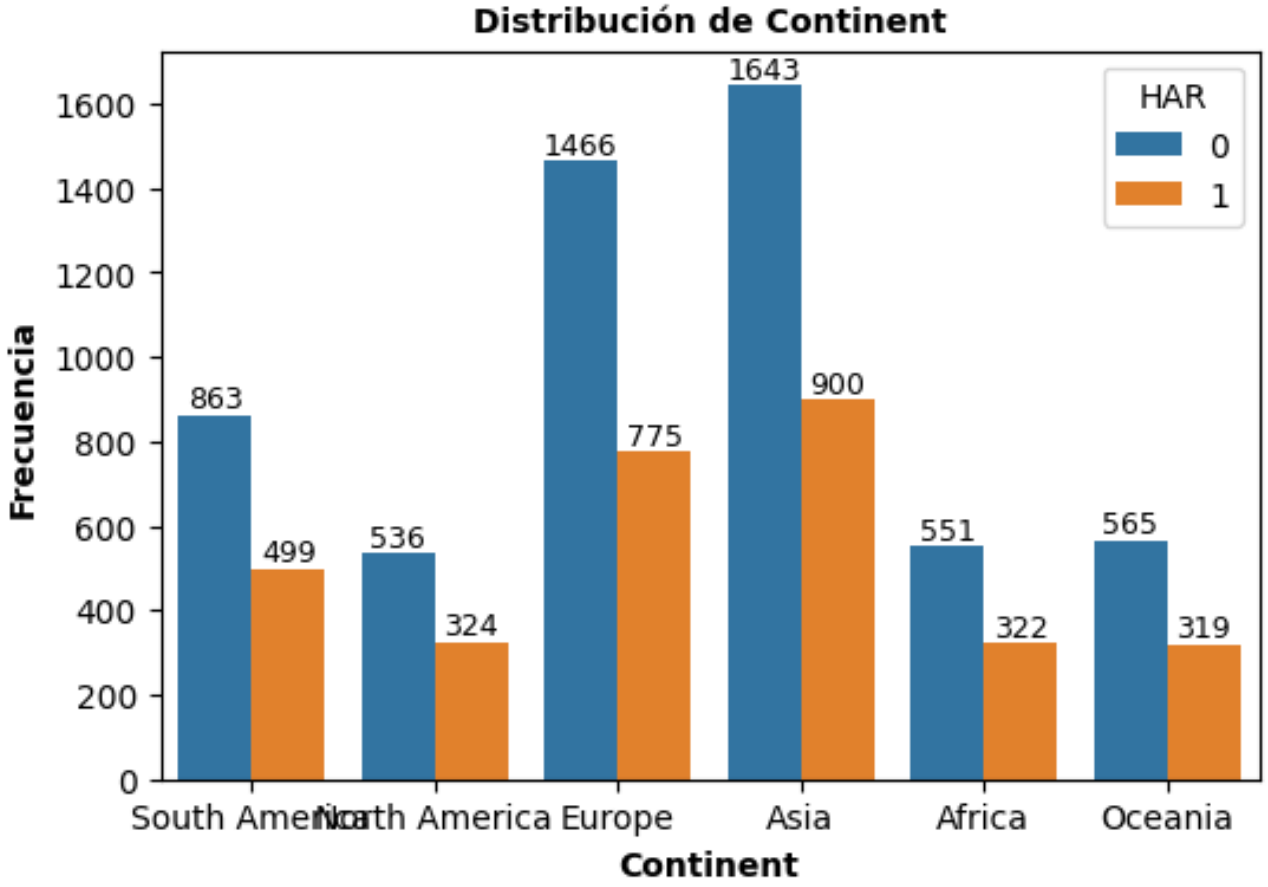
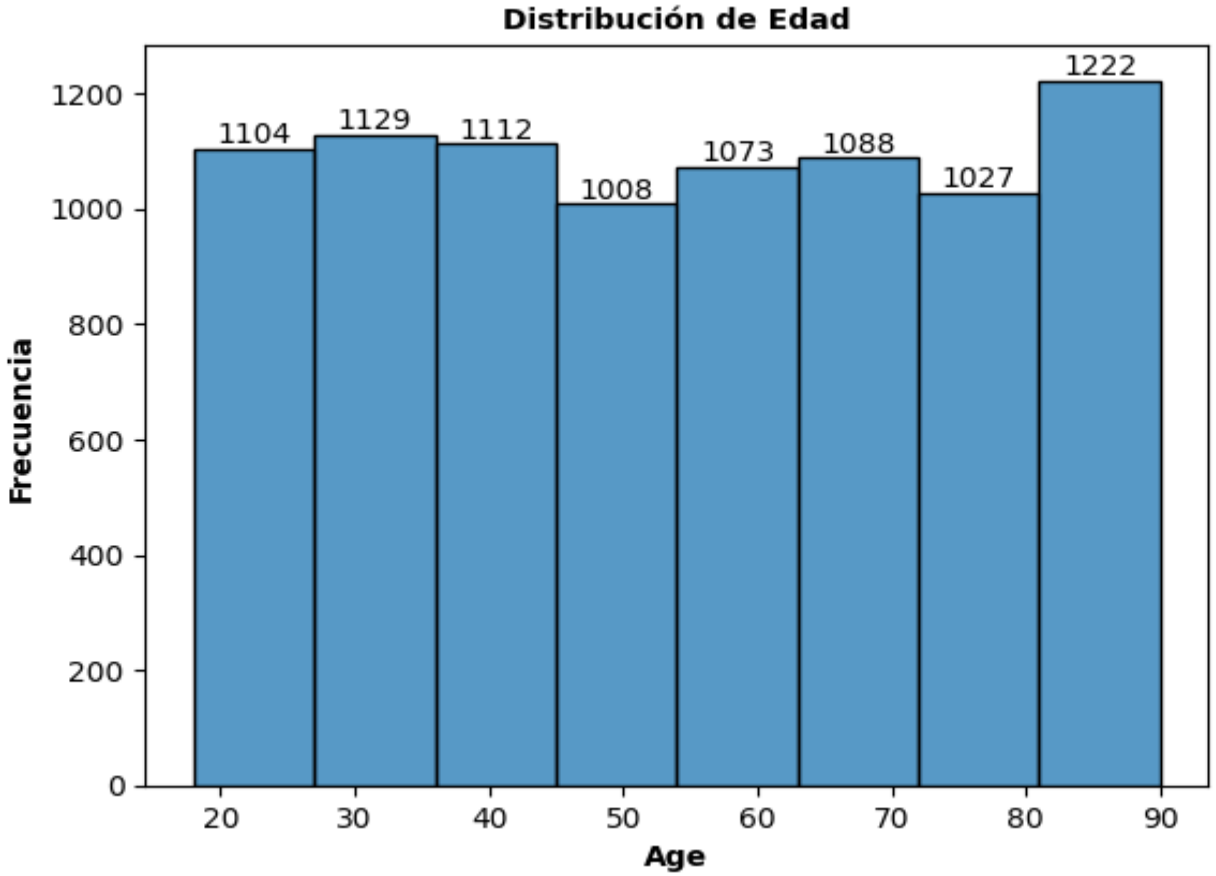


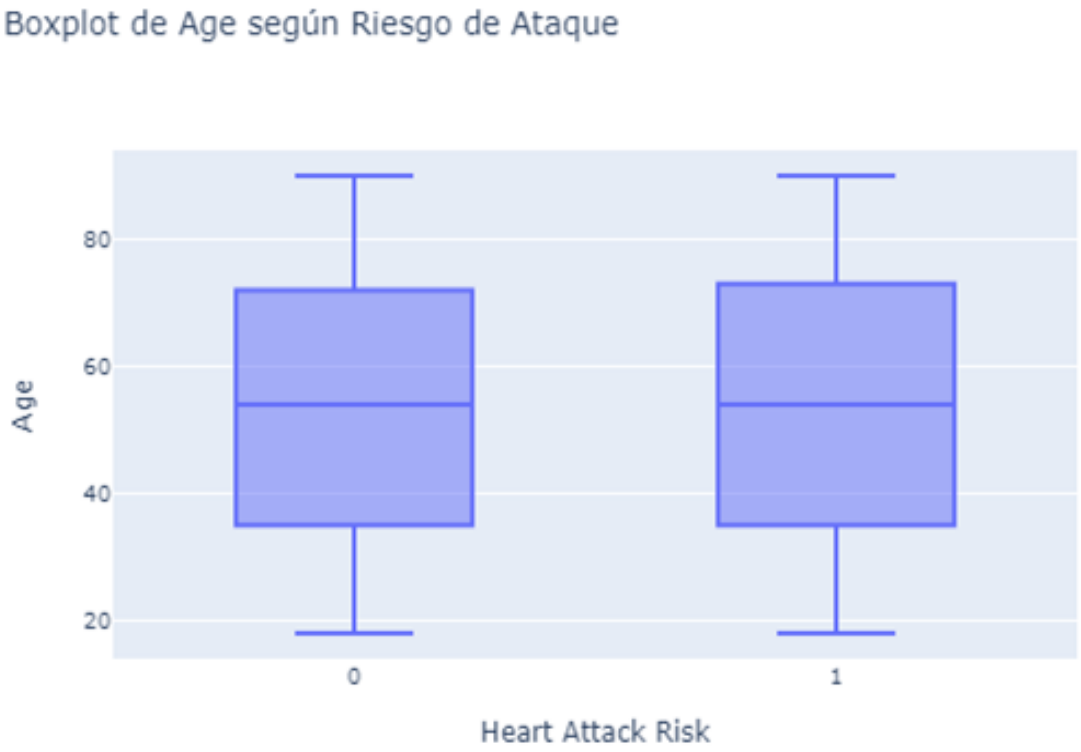
Tabla de porcentajes por Continente:

% Cont_Africa: 36.88
% Cont_Asia: 35.39
% Cont_Europe: **34.58**
% Cont_North America: **37.67**
% Cont_Oceania: 36.09
% Cont_South America: 36.64

La relación respecto a la variable target tiene un comportamiento muy similar, el porcentaje de casos con riesgo cardíaco en cada continente va desde un 34.58% a un 37.67%

¿Cuál es la relación entre la **Edad y el Riesgo de Ataque Cardíaco?**





La variable Edad representa una distribución uniforme, no presenta 'colas' o 'picos' en su visualización; podemos hablar de una distribución de valores 'similar' entre los intervalos.

INSIGHTS & **RECOMENDACIONES**

INSIGHTS & **PASOS A SEGUIR**

Insights A PARTIR DE VISUALIZACIONES Y EXPLORACIÓN DE DATOS

- **Variable Physical Activity Days Per Week,** presenta una posible variabilidad en su estudio que la puede indicar como un factor de influencia. Debemos considerarla a futuro en nuestros análisis multivariados -en particular con Continente-.
- **Variable Triglycerides,** comportamiento similar a Días de actividad física.
- **Variable Systolic_pressure,** comportamiento similar a Días de actividad física.

Insights A PARTIR DE INFORME SEGÚN PANDAS PROFILING

- **Variable Smoking,** presenta una correlación con sexo y edad. Esto podría significar en un análisis multivariado un posible factor de riesgo que en la combinación de factores tenga una mayor incidencia en nuestra predicción.
- **Variable Continent -con one hot encode aplicado,** se destaca en varios casos su desbalanceo pero el mismo es explicado a partir de la aplicación del método.

Insights A PARTIR DE SELECCIÓN DE VARIABLES POR MÉTODOS DE ML

- **Variables Diabetes, Systolic_pressure, Cholesterol y Sleep Hours per day,** son las resultantes a partir de la iteración aplicada con los métodos de feature selection. En la notebook trabajada encontraremos un summary para entender su incidencia estadística en el modelo de regresión logística planteado.

Pasos a Seguir:

- Aplicar modelos de mayor complejidad que nos permitan establecer a partir de un análisis multivariado la posibilidad de encontrar factores de mayor incidencia en nuestro dataset.
- Explorar cómo determinar una mejor configuración de nuestros modelos a partir de hiper parámetros.