

Análisis de Datos (Bivalente): Numéricas. Correlaciones

Vamos a analizar las variables numéricas con el diagrama de dispersión y la correlación matemática. Eso es así por la naturaleza de nuestros datos, pero si nuestro dataset fuera de cotizaciones de empresas en los últimos 365 días entonces añadiríamos a lo anterior (de hecho, lo haríamos lo primero) un gráfico de evolución temporal de las mismas en el que ambas estuvieran presentes.

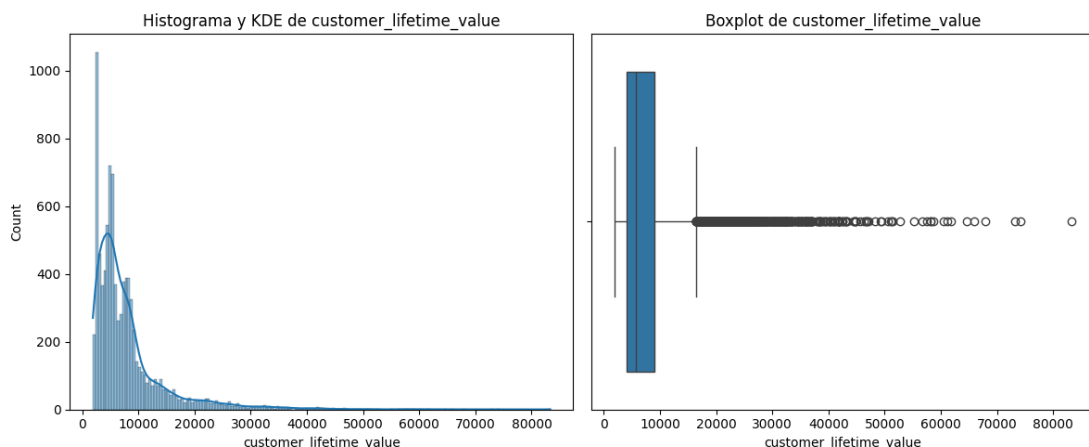
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
from bootcampviztools import plot_categorical_numerical_relationship, plot_combined_graphs, \
    pinta_distribucion_categoricas, plot_grouped_boxplots, grafico_dispersion_con_correlacion
```

```
df_seguros = pd.read_csv("./data/Marketing-Customer-Analysis.csv")
df_air_jun = pd.read_csv("./data/dataset_viajes_jun23.csv")
```

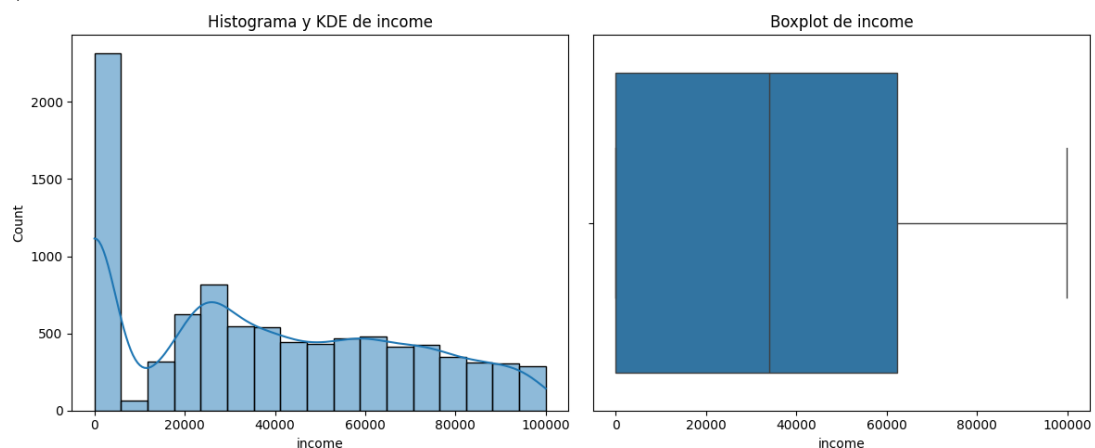
Caso 1. Seguros: Bivalente CLV-Ingresos

#1 Tener la referencia de la **distribución de valores por separado de cada variable**

```
plot_combined_graphs(df_seguros,["customer_lifetime_value"])
(2,)
```



```
plot_combined_graphs(df_seguros,["income"])
(2,)
```



#2 Como ambas variables están medidas en la misma unidad mostramos una comparación numérica de sus describe:

```
df_seguros[["customer_lifetime_value","income"]].describe()
```

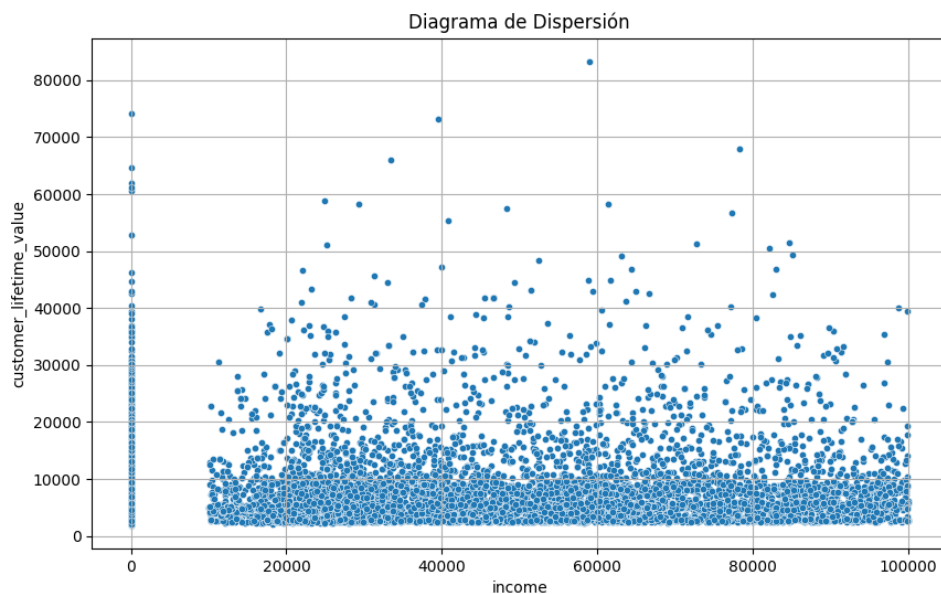
[4]:	customer_lifetime_value	income
count	9134.000000	9134.000000
mean	8004.940475	37657.380009
std	6870.967608	30379.904734
min	1898.007675	0.000000
25%	3994.251794	0.000000
50%	5780.182197	33889.500000
75%	8962.167041	62320.000000
max	83325.381190	99981.000000

#3 Y comparamos los datos unos con otros. Utilizamos un diagrama de dispersión *scattered plot*

Un scatter plot, o diagrama de dispersión, es un tipo de gráfico en el que cada punto representa una observación individual, con la posición en el eje horizontal (X) indicando el valor de una variable y la posición en el eje vertical (Y) indicando el valor de la otra variable.

Los scatter plots son útiles para observar la relación entre dos variables numéricas. Se utilizan para detectar correlaciones, tendencias, agrupaciones, o para identificar patrones inusuales o valores atípicos (outliers) en los datos.

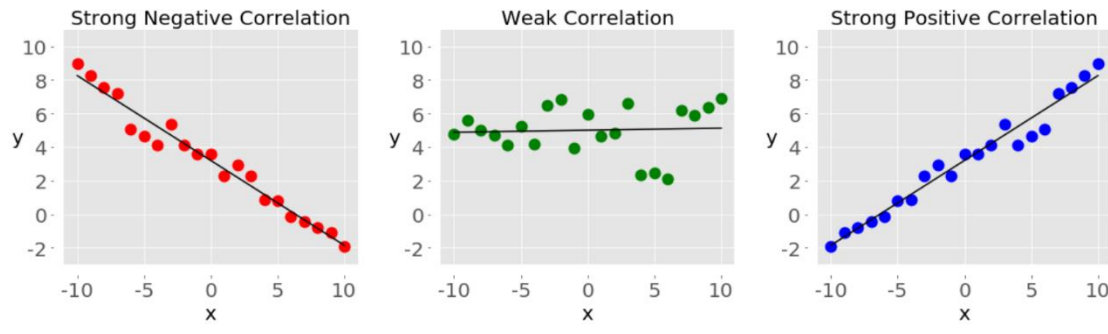
```
grafico_dispersion_con_correlacion(df_seguros,"income","customer_lifetime_value", 20)
```



Correlación y relación entre variables numéricas

La **correlación** trata de establecer la relación o **dependencia que existe entre las dos variables**. Es decir, determinar si los cambios en una de las variables influyen en los cambios de la otra. En caso de que suceda, diremos que las variables están correlacionadas o que hay correlación entre ellas. La correlación es **positiva** cuando los valores de las variables aumentan juntos; y es **negativa** cuando un valor de una variable se reduce cuando el valor de la otra variable aumenta.

De primeras se suele analizar usando un diagrama de dispersión como el mostrado y los ejemplos prototípicos son:



[En el primero se ve que al aumentar la variable x disminuye la y, se dice que hay una fuerte correlación negativa, en el segundo los cambios en x, no parecen afectar a los cambios en y, se dice que no hay correlación o que esta es muy débil, y finalmente cuando los cambios de x se relacionan con cambios de y en el mismo sentido se habla de una correlación fuerte]

¿A cuál de los tres se te parece el gráfico de dispersión de ingresos y CLV? A los dos de los extremos no, eso seguro. Por eso podemos pensar que no hay correlación entre ambos, es decir el CLV varía de forma similar y en rangos similares independientemente del income del cliente.

Pero no acaba aquí la cosa hay un coeficiente matemático (de hecho hay varios) que miden esa relación. El más utilizado, y el que vamos a usar nosotros es el **coeficiente o correlación de Pearson**. Cuya fórmula te dejo aquí (no te asustes que usaremos funciones y métodos para obtenerlo):

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

En esta fórmula:

- r es el coeficiente de correlación de Pearson.
- x_i y y_i son los valores individuales de las dos variables.
- \bar{x} y \bar{y} son las medias de las variables x y y , respectivamente.
- n es el número de observaciones (parejas de valores).

El coeficiente varía entre -1 y 1, donde_

- 1 significa una correlación positiva perfecta
- -1 una correlación negativa perfecta

- 0 indica que no hay correlación **lineal** entre las variables. [marco lineal en negrita porque puede haber otro tipo de relación]

Y yendo a lo práctico, tenemos varias formas de obtener ese coeficiente:

```
df_seguros[["income","customer_lifetime_value"]].corr()
```

```
[6]:
```

	income	customer_lifetime_value
income	1.000000	0.024366
customer_lifetime_value	0.024366	1.000000

```
from scipy.stats import pearsonr
```

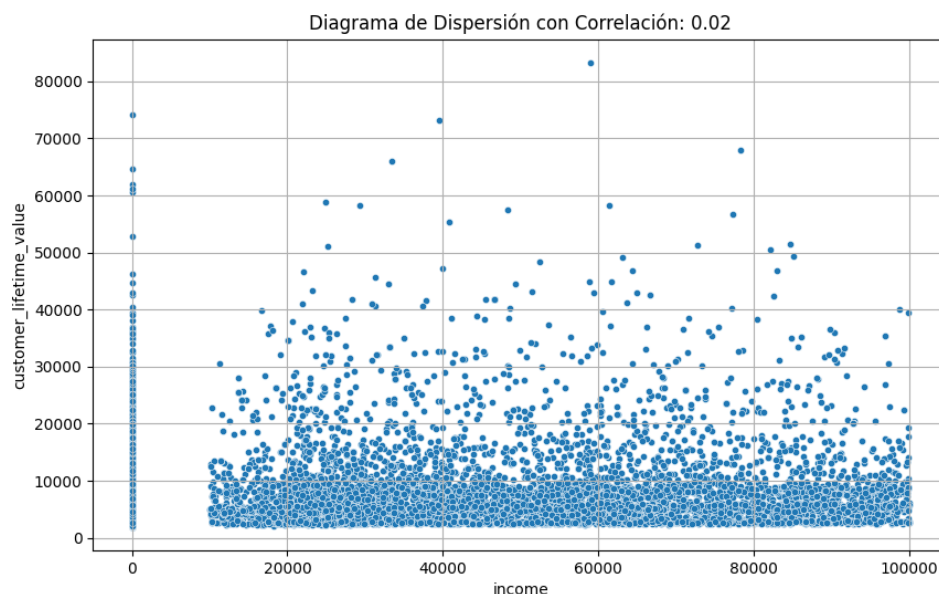
```
pearsonr(df_seguros["income"], df_seguros["customer_lifetime_value"], alternative= "less")
```

[PearsonRResult\(statistic=0.024365660747409063, pvalue=0.9900625445226664\)](#)

De ese pvalue hablaremos en breve.

```
grafico_dispersion_con_correlacion(df_seguros,"income","customer_lifetime_value",
mostrar_correlacion = True)
```

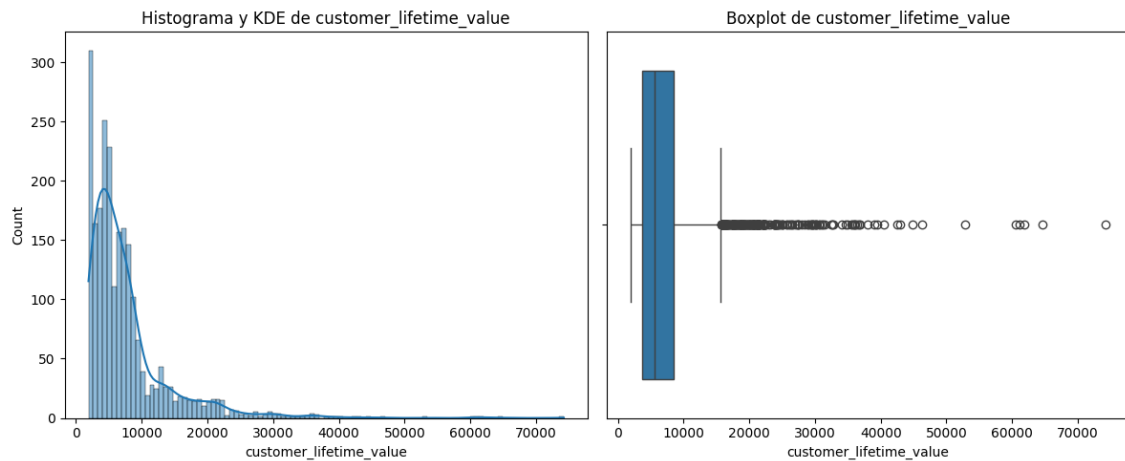
20,



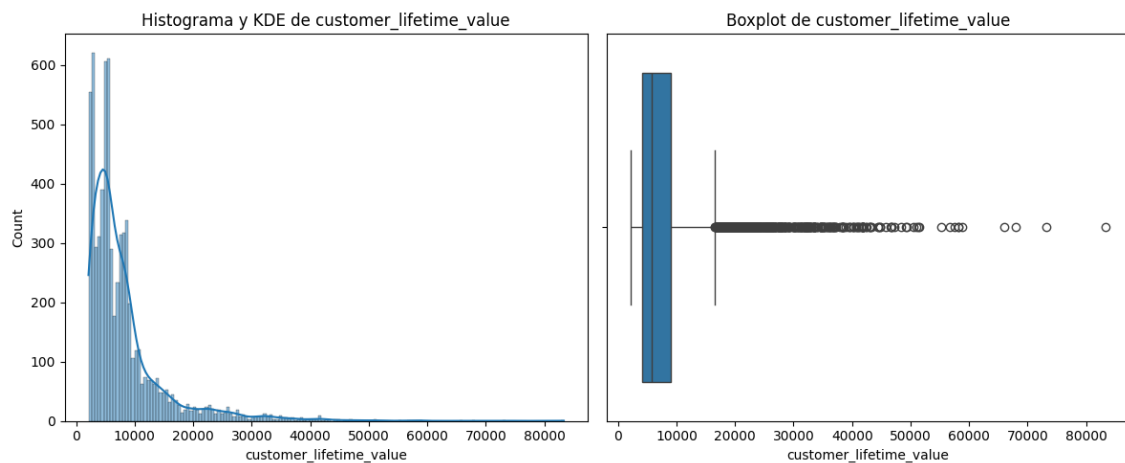
Básicamente: No están correlacionadas linealmente, lo cual ya nos lo podemos apuntar como un elemento más de nuestro análisis.

Pero no termina aquí el análisis, ahora está claro que es necesario hacer el análisis en dos. Fijate en el salto entre 0 y el siguiente valor. **El 0 es un outlier de Income** que no habíamos visto antes. Así que hay que quitarlo y quizás analizar por separado y preguntar qué tipo de cliente es este (seguros contratados por otros?) y también así dejar limpiar la variable income para analizarla con el resto.

```
plot_combined_graphs(df_seguros.loc[df_seguros.income == 0], ["customer_lifetime_value"])
(2,)
```



```
plot_combined_graphs(df_seguros.loc[df_seguros.income > 0], ["customer_lifetime_value"])
(2,)
```



Tiene pinta de que income = 0 es una forma de marcar "No sabemos el income" del cliente. Esto es otro dato a preguntar, y claramente ahora tendría que hacer los dos análisis por separado.

Contraste de hipótesis y el p-value (una intro muy grosera a modo de adelanto)

El contraste de hipótesis es básicamente una prueba para dar "credibilidad matemática" a afirmaciones como "la variable x e y están correlacionadas". Por ejemplo, si dos variables muestran un coeficiente de 0.02 de Pearson, ¿cuánto podemos asegurar de que esto implica que son independientes (linealmente)?

A eso nos responde "el contraste de hipótesis" (ver en estadística inferencial): t de student, chi cuadrado, test de Wilcoxon, test de Shapiro-Wilk o prueba F de Fisher.

En esos test planteamos nuestra hipótesis y obtenemos un valor denominado p-value que es básicamente la probabilidad de que la hipótesis contraria a la tuya sea lo normal. Es decir, cuanto más bajo ese p más "seguro" podrías estar de tu suposición. Si ese valor p está por debajo de un umbral (típicamente 0.05) entonces pensaremos que nuestra suposición (que dos variables tienen relación) puede ser cierta con bastante confianza. En términos formales nos dirá que la relación o falta de ella es estadísticamente significativa.