**Análisis de Datos (Bivariante): Combinada categórica-numérica**

**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, plot\_grouped\_histograms *# Este módulo lo tienes en el directorio de los notebooks, puede abrirlo y jugar con las funciones que contiene*

df\_seguros **=** pd.read\_csv("./data/Marketing-Customer-Analysis.csv")

df\_air\_jun **=** pd.read\_csv("./data/dataset\_viajes\_jun23.csv")

**Caso 1. Seguros: Bivariante response-CLV**

Pues no nos queda otra que analizar la relación de las variables target o directoras entre sí.

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

pinta\_distribucion\_categoricas(df\_seguros,["response"], relativa**=** **True**, mostrar\_valores**=** **True**)

Gráfico, Gráfico en cascada

Descripción generada automáticamente

plot\_combined\_graphs(df\_seguros,["customer\_lifetime\_value"])

(2,)

Gráfico

Descripción generada automáticamente

**#2** Mostramos diferentes medidas de la variable numérica en función de los valores de la categórica. **Empezamos con la mediana:**

plot\_categorical\_numerical\_relationship(df\_seguros,"response","customer\_lifetime\_value", measure **=** "median")

Gráfico, Gráfico de barras

Descripción generada automáticamente

Aquí hay una ligera diferencia, pero con ver la gráfica de histograma-densidad o el boxplot de CLV ya sabemos que la tendencia central no es la mejor representación.

**#3** Mostramos diferentes medidas de la variable numérica en función de los valores de la categórica. **Comparemos los diagramas de caja:**

plot\_grouped\_boxplots(df\_seguros,"response","customer\_lifetime\_value")

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

**#4** Mostramos diferentes medidas de la variable numérica en función de los valores de la categórica. **Comparemos los histogramas:**

plot\_grouped\_histograms(df\_seguros,"response", "customer\_lifetime\_value",2)

Gráfico, Histograma

Descripción generada automáticamente con confianza media

Tiene pinta de que salvo por el hecho de que los rangos no coinciden, no hay dependencia entre ambas. Es decir que la respuesta comercial es independiente del CLV. Este resultado me invita a:

* Hacer uso de los test estadísticos para comprobar la independencia de las dos variables [Eso del test lo veremos en la parte de estadística inferencial]
* Hacer el análisis descartando outliers, o mejor, hacer un análisis sin outliers y otro para outliers.

En principio me quedo con la intuición/hipótesis de que el CLV y responder a las ofertas de marketing es aparentemente independiente.

Y así seguiría con el resto de las variables numéricas y response, por ejemplo. Pero también tendría que hacerlo con las otras categóricas y CLV, porque de hecho es lo que nos han pedido.

**Ejemplo extra: CLV-Education**

pinta\_distribucion\_categoricas(df\_seguros,["education"], relativa**=** **True**, mostrar\_valores**=** **True**)

Gráfico, Gráfico en cascada

Descripción generada automáticamente

plot\_combined\_graphs(df\_seguros,["customer\_lifetime\_value"])

(2,)

Gráfico

Descripción generada automáticamente

plot\_categorical\_numerical\_relationship(df\_seguros,"education","customer\_lifetime\_value", measure **=** "median")

Gráfico, Gráfico de barras

Descripción generada automáticamente

Este es un dato curioso, si la mediana es buena representación, no correla linealmente con el grado de estudios (que es una categórica que podríamos considerar ordinal), de hecho, los doctores "pagan" menos o eso parece (por ahora eso parece, tendría aquí una hipótesis curiosa para seguir)

plot\_grouped\_boxplots(df\_seguros,"education","customer\_lifetime\_value")

Gráfico

Descripción generada automáticamente

Así es a ojo y en este momento sacaría el describe agrupado de CLV para ver los percentiles, pero tiene pinta de que los que tienen master son los de mayor percentil 75 y que los Doctor tienen mayor dispersión. Esto no me sirve de mucho. Sigamos con las distribuciones. High School or Below parece más compacto y con mejores números que el resto. Fíjate donde están los mayores outliers de CLV, eso me plantearía estudiarlos con cuidado, porque no me esperaba que fueran los que más CLV tuviesen. Lo cual no tiene por qué ser nada raro (que sean ellos, pueden ser clientes con flotas de coches, o autónomos con varios coches) pero despierta mi curiosidad.

plot\_grouped\_histograms(df\_seguros,"education", "customer\_lifetime\_value", 5)

Gráfico, Histograma

Descripción generada automáticamente

Aunque al comparar las distribuciones entre sí, lo cierto es que no parece haber tanta diferencia, siguen perfiles similares, mucho, pero es cierto que los doctores son muchos menos y eso crea ese efecto. En este caso sí que quería ver números:

df\_seguros.groupby("education")["customer\_lifetime\_value"].describe()

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente con confianza media

Sin duda, repetiría este análisis separando los outliers para confirmar que el nivel de educación correla inversamente con el CLV aunque sea muy ligeramente. Y otra métrica interesante que me gustaría comprobar es el porcentaje de outliers en total y por grupo.

**Caso 2. Viajes: Bivariante aircompany-ingresos**

pinta\_distribucion\_categoricas(df\_air\_jun,["aircompany"], relativa**=** **True**, mostrar\_valores**=** **True**)

Gráfico, Gráfico de barras

Descripción generada automáticamente

plot\_combined\_graphs(df\_air\_jun,["ingresos"])

(2,)

Gráfico, Histograma

Descripción generada automáticamente

plot\_categorical\_numerical\_relationship(df\_air\_jun, "aircompany","ingresos",measure**=**"median")

*# Utilizando mediana*

Gráfico, Gráfico de barras

Descripción generada automáticamente

Este resultado parece bastante potente. Tenemos el mayor número de vuelos, pero nuestros ingresos considerando la mediana son los menores de todos. Ojo, esto sólo es para quedárselo apuntado, pero es ya una hipótesis de trabajo potente, ¿es esta una mala foto? Puede que seamos la que menos ingresa, pero la que menos consume y al final el beneficio se equilibre o sea mayor (es cierto que cada viaje tiene un coste no imputado...) En general hemos encontrado otro punto donde excavar... y sacarle todo el jugo, vamos a confirmarlo viendo el resto de comparativas

plot\_grouped\_boxplots(df\_air\_jun,"aircompany","ingresos")

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Esto tiene mala pinta sin duda, para Tabarair, para nuestro estudio es magnífico porque nos abre todo un camino de trabajo.

plot\_grouped\_histograms(df\_air\_jun,"aircompany", "ingresos", 5)

Gráfico, Histograma

Descripción generada automáticamente

plot\_grouped\_histograms(df\_air\_jun,"aircompany", "ingresos", 2)

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Se ve claramente que TabarAir tiene el menor rango de ingresos, que ya lo sabíamos de ver los boxplots, y que casi los concentramos, habrá que compararlo con distancias y consumos para chequear la eficiencia. Y sin ser un experto parece que Pampangea y Moldaviair tienen una dispersión de ingresos que invita a pensar que están mejor diversificados (tienen vuelos de todos los gustos, aunque esto es otra hipótesis para confirmar). Lo importante: Tabarair es la empresa con más vuelos, menos ingresos en mediana, menor rango de ingresos con un pico en vuelos de pocos ingresos y una distribución casi gaussiana. Mientras Airnar con menos vuelos es la tercera en ingresos.