**Análisis de Datos (Bivariante): Test de relación entre categórica y numérica**

Para ver si existe una relación estadística significativa entre una variable categórica y una numérica podemos aplicar varios test en función de la cardinalidad de la categórica y de otras condiciones.

**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\_categorical\_relationship\_fin, plot\_grouped\_histograms

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

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

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

Como siempre primero los valores por separados y el análisis visual al combinarlas.

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

Gráfico, Gráfico de barras

Descripción generada automáticamente con confianza media

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

(2,)

Gráfico

Descripción generada automáticamente

plot\_grouped\_histograms(df\_seguros, cat\_col**=**"response",num\_col**=**"customer\_lifetime\_value", group\_size**=** 2)

Gráfico, Histograma

Descripción generada automáticamente con confianza media

Como puedes ver, response es una variable binaria, en esa situación se suele utilizar el test t-Student si se cumplen determinadas condiciones (entre otras que los valores estén distribuidos como una normal, es decir una campana de gauss) pero que no se cumplen así que nosotros usaremos [la prueba U de Mann-Whitney](https://datatab.es/tutorial/mann-whitney-u-test):

**Aplicando la prueba U de Mann-Whitney:**

**#0 Importar el test de scipy**

**from** scipy.stats **import** mannwhitneyu *# Sí, de scipy.stats :-)*

**#1 Separar los datos en dos grupos según la variable categórica**

grupo\_a **=** df\_seguros.loc[df\_seguros.response **==** "Yes"]["customer\_lifetime\_value"]

grupo\_b **=** df\_seguros.loc[df\_seguros.response **==** "No"]["customer\_lifetime\_value"]

**#2 Aplicar la prueba y mostrar los resultados**

u\_stat, p\_valor **=** mannwhitneyu(grupo\_a, grupo\_b)

print("Estadístico U:", u\_stat)

print("Valor p:", p\_valor)

Estadístico U: 5060988.0

Valor p: 0.5168872245291198

**#3 Interpretación del valor p (p-value)**

En este caso la hipótesis nula (o directora) es que no hay diferencia estadística significativa entre las medianas de los valores, si hubiéramos obtenido un p-value por debajo de 0.05, hubiéramos rechazado la hipótesis nula 🡪 hay confianza estadística en que las medianas de cada grupo son diferentes para cada valor y que hay una relación entre los valores de las dos variables. En este caso, no podemos rechazar la hipótesis, no podemos decir que haya esa relación entre las variables.

En resumen: Según este test, no podemos asegurar que la respuesta y el CLV tengan relación. Lo que reafirma nuestra confirmación visual, pero no confirma.

**Caso 2. Viajes: Relación entre aircompany e ingresos**

Pintamos los valores:

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

Gráfico, Gráfico de barras

Descripción generada automáticamente

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

(2,)

Gráfico, Histograma

Descripción generada automáticamente

plot\_grouped\_histograms(df\_air\_jun, "aircompany","ingresos",group\_size**=** 2)

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

En este caso la variable categórica no es binaria y no podemos usar la prueba U. Para estos casos empleamos el método ANOVA.

**Aplicando ANOVA:**

Nota: ANOVA es toda una metodología de análisis de varianza con categóricas y numéricas, Tiene también sus suposiciones de partida que puede que no estemos cumpliendo (pero se escapa del ámbito del bootcamp cubrir este punto). Aquí encontrarás más información y detalle sobre [Anova](https://datatab.es/tutorial/anova" \t "_blank)

**#1 Importar librerias**

**from** scipy **import** stats

**#2 Separar en grupos según los valores de las categorías**

grupos **=** df\_air\_jun['aircompany'].unique() *# Obtener los valores únicos de la columna categórica, en este caso la compañía área*

ingresos\_por\_company **=** [df\_air\_jun[df\_air\_jun['aircompany'] **==** grupo]['ingresos'] **for** grupo **in** grupos] *# obtenemos los ingresos por compañía y los incluimos en una lista*

ingresos\_por\_company

[0 14232.65

9 110108.07

17 518685.45

21 799063.90

23 432558.22

...

573 732878.07

577 754191.00

580 928050.52

596 87307.63

597 339943.47

Name: ingresos, Length: 100, dtype: float64,

1 468527.19

33 273824.24

36 235350.43

40 25642.94

41 417965.59

...

575 429377.42

581 1004810.48

589 398905.47

590 31227.48

598 752212.98

Name: ingresos, Length: 108, dtype: float64,

2 584789.19

4 438535.07

10 111056.67

12 184079.01

13 46200.30

...

572 258737.92

582 423371.06

587 404542.35

593 550268.31

599 330320.81

Name: ingresos, Length: 142, dtype: float64,

3 233342.51

5 728045.68

6 13805.52

11 764998.83

16 719577.55

...

578 12372.41

579 409713.51

583 1006880.11

584 70674.57

585 225493.20

Name: ingresos, Length: 133, dtype: float64,

7 1056735.47

8 600836.96

18 584677.98

19 245091.83

26 94840.33

...

588 20749.29

591 952688.76

592 583465.41

594 28007.41

595 442668.52

Name: ingresos, Length: 117, dtype: float64]

**#3 Aplicamos el ANOVA**

f\_val, p\_val **=** stats.f\_oneway(**\***ingresos\_por\_company) *# El método \*lo que hace es separar todos los elementos de la lista y pasárselos como argumento a la función*

print("Valor F:", f\_val)

print("Valor p:", p\_val)

Valor F: 15.68687291880937

Valor p: 3.267745855663723e-12

**#4 Interpretación del ANOVA**

En este caso la hipótesis nula vuelve a ser que las distribuciones de valores para cada compañía no tienen relación estadística, pero ahora como el p-value es menor que 0.05, rechazamos la hipótesis nula y reafirmamos la alternativa o contraria, es decir que sí que los ingresos no se parecen para cada compañía y que podemos extrapolar ese valor a otro conjunto de compañías con datos similares.

**RESUMEN PASOS ANALISIS BIVARIANTE**

1. Establece una lista de combinaciones que quieras analizar sí o sí: Básate en las preguntas e hipótesis de partida que tengas, en la prioridad de las variables y en los resultados y notas de tu análisis univariante.
2. Aplica a cada combinación lo que has visto en el workout. Ahora ya empezarán a salir mensajes y conclusiones que sí pueden aparecer en el informe y presentación finales. Puede que surjan otras hipótesis y tengas que volver atrás.
   * Si tienes muestras y quieres extrapolar resultados, acuérdate de aplicar el contraste de hipótesis.
3. Haz una lista de preguntas que puedas responder ya y busca esa respuesta (ya no va de analizar sistemáticamente sino de filtrar, operar, agrupar los dataframes hasta que encuentres esa respuesta). Por ejemplo: No hace falta esperar para ver con más detalle qué ocurre con las ofertas de tipo 2 las que tenían más éxito que el resto de las ofertas.
4. Lánzate a analizar varias variables a la vez (más de 2) siempre guiado por las preguntas e hipótesis.