

Fastbook 06

Estadística Aplicada al Marketing

Principales técnicas estadísticas



06. Principales técnicas estadísticas

Con el fastbook anterior nos adentramos en el mundo de la estadística:

- **Población y muestra.** Nos familiarizamos con los conceptos básicos de estadística en los que se fundamentan todos los estudios: población y muestra. Conocemos también el significado de tamaño poblacional o muestral y los distintos tipos de muestreo.
- **Tipos de variables.** Descubrimos los tipos de variables que existen y cómo se suele hacer referencia a ellos desde el punto de vista de bases de datos o programación.
- **Medidas de centralización.** Comprendemos el significado de medidas de tendencia central, haciendo hincapié, sobre todo, en las 3 principales: media (media ponderada), mediana y moda. Aprendimos a calcularlas y los beneficios, limitaciones, desventajas que tiene cada una.
- **Medidas de dispersión.** Descubrimos el significado de medidas de dispersión, y el significado y cálculo de las dos medidas principales: varianza y desviación típica.
- **Medidas de posición.** Por último, conocimos las principales medidas de posición, la división que hacen de la muestra y su relación con la mediana. Las medidas que estudiamos fueron los cuartiles, deciles y percentiles.

Con este fastbook complementaremos las medidas anteriores junto con las tablas de frecuencia, los outliers, covarianza, correlaciones y análisis exploratorio. La combinación de todas estas técnicas es la base de cualquier análisis descriptivo, el que sirve para tener una visión completa de la distribución de los datos.

Autora: Patricia Martín González

Tablas de frecuencia

Outliers

Covarianza

Correlación

Análisis exploratorio de datos (EDA)

Resumen

Tablas de frecuencia

X Edix Educación

A diferencia de las medidas vistas anteriormente, las **tablas de frecuencia** (o tablas de contingencia) se suelen usar para **medir la distribución de una variable numérica a lo largo de una variable categórica** (o transformaciones de numéricas en categóricas).

Las tablas de frecuencia (o tablas de contingencia) son un recurso muy utilizado y práctico para presentar resultados de variables numéricas y no numéricas. Permiten entender los datos y ver su distribución de forma muy simple mediante cálculos aritméticos básicos (conteos, sumas, porcentajes, promedios, etc.).

Supongamos que queremos medir la lealtad de 50 de nuestros clientes en relación a nuestra marca:

¿En qué medida recomendarías a un amigo nuestra marca? Valora de 0 a 10, donde 0 es nada probable y 10 extremadamente probable.

Los resultados que obtenemos son:

Individuo	id1	id2	id3	id4	id5	id6	id7	id8	id9	id10	id11	id12	id13	id14
Valoración	9	10	5	7	8	7	6	7	8	0	8	0	9	10

Individuo	id15	id16	id17	id18	id19	id20	id21	id22	id23	id24	id25	id26
Valoración	9	8	5	8	8	9	7	8	7	9	8	9
Individuo	id27	id28	id29	id30	id31	id32	id33	id34	id35	id36	id37	id38
Valoración	7	4	9	9	7	4	10	6	8	9	10	9

Individuo	id39	id40	id41	id42	id43	id44	id45	id46	id47	id48	id49	id50
Valoración	10	2	6	7	10	3	9	8	5	7	8	5

Una vez que tenemos los datos, creamos la tabla de frecuencia.

La primera columna describe los posibles valores de la variable y suele representarse como X_i .

La segunda recibe el nombre de **frecuencia absoluta** (n_i) y representa el número de veces que se repiten los valores X_i de la variable de estudio.

La suma de las frecuencias de todos valores es el tamaño poblacional (N).

$$\sum_{i=1}^k n_i = n_1 + n_2 + \dots + n_k = N$$

La tabla de frecuencia en nuestro caso sería:

Valoración (X_i)	Frecuencia (n_i)
0	2
1	0
2	1
3	1
4	2
5	4
6	3
7	9
8	11
9	11
10	6
Total (N)	50

Como vemos en la tabla, las valoraciones más frecuentes de los encuestados han sido 8 y 9, con 11 respuestas cada una. Los resultados sobre satisfacción del cliente a priori parecen buenos.

Sobre la tabla base, vamos el resto de métricas (aunque de forma teórica pueda parecer algo complicado, con el ejemplo veremos que es muy sencillo):

La frecuencia absoluta acumulada (N_i)

La frecuencia absoluta acumulada (N_i) de un valor X_i es la suma de las frecuencias absolutas (n_i) de los valores menores o iguales que X_i .

$$N_i = n_1 + n_2 + \dots + n_{i-1} + n_i$$

La frecuencia relativa (f_i)

La frecuencia relativa (f_i) de un valor X_i es la frecuencia relativa dividida por el número total de elementos N. Son valores entre 0 y 1, y suelen leerse como probabilidad, es decir, proporción de casos frente al total.

$$f_i = \frac{n_i}{N}$$

Suma de las frecuencias relativas

La suma de las frecuencias relativas debe ser 1.

$$\sum_{i=1}^k f_i = f_1 + f_2 + \dots + f_k = 1$$

La frecuencia relativa acumulada (F_i)

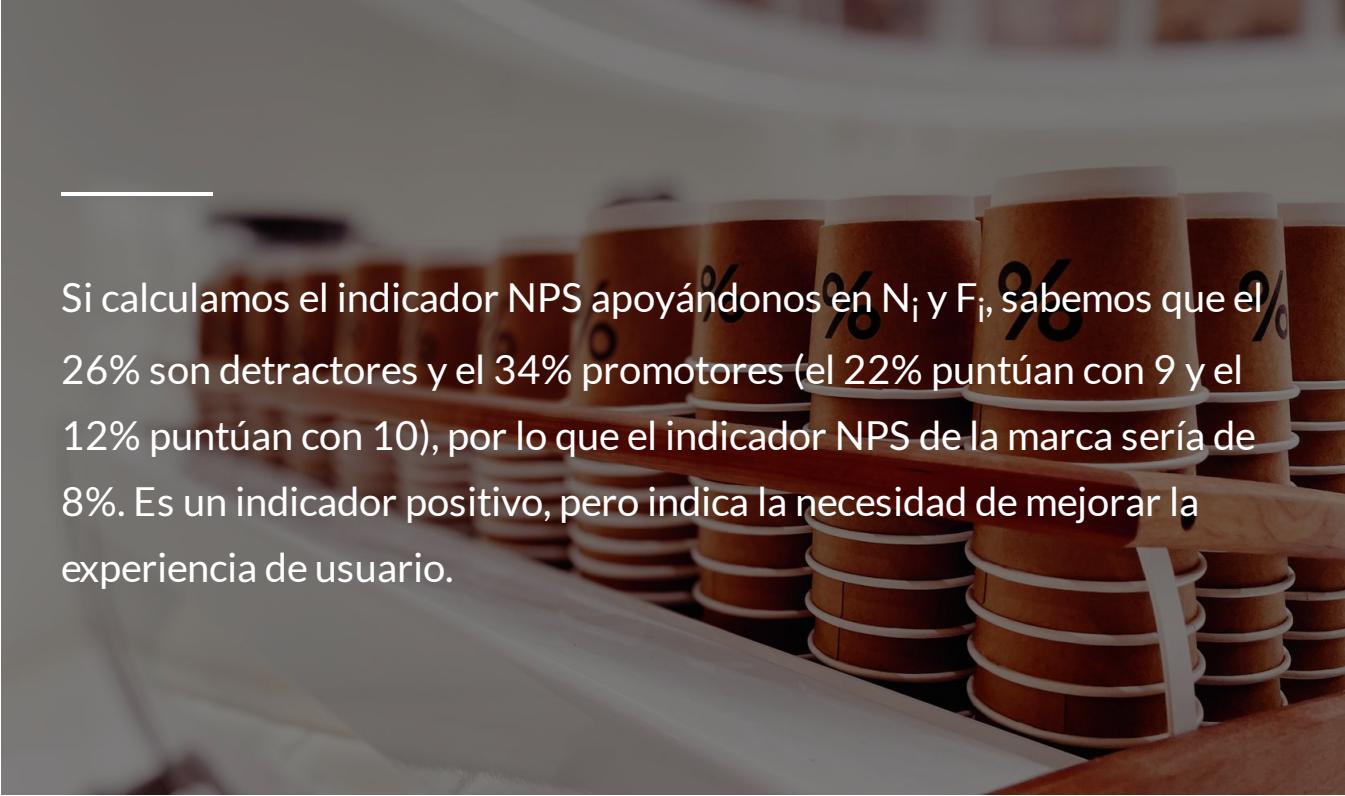
La frecuencia relativa acumulada (F_i) de un valor X_i es la suma de las frecuencias relativas de los valores menores o iguales que X_i , o la frecuencia absoluta acumulada dividida por el número total de individuos (N).

$$F_i = \frac{N_i}{N} = f_1 + f_2 + \dots + f_i$$

La tabla de frecuencias con estas métricas agregadas sería:

Valoración (X)	Frecuencia absoluta (n)	Frecuencia absoluta acumulada (N _i)	Frecuencia relativa en porcentaje (f _i)	Frecuencia relativa acumulada (F _i)
0	2	2	4%	4%
1	0	2	0	4%
2	1	3	2%	6%
3	1	4	2%	8%
4	2	6	4%	12%
5	4	10	8%	20%
6	3	13	6%	26%
7	9	22	18%	44%
8	11	33	22%	66%
9	11	44	22%	88%
10	6	50	12%	100%
Total	50 (N)		100%	

A la vista de los resultados, cabe destacar que tan solo 13 personas (detractores) respondieron con una puntuación de 6 o inferior (frecuencia absoluta acumulada N_i), que suponen el 26% de la población (frecuencia relativa acumulada F_i). También llama la atención que entre las puntuaciones de 8 y 9 encontramos el 44% de las valoraciones (frecuencia relativa en porcentaje f_i).



Si calculamos el indicador NPS apoyándonos en N_i y F_i , sabemos que el 26% son detractores y el 34% promotores (el 22% puntúan con 9 y el 12% puntúan con 10), por lo que el indicador NPS de la marca sería de 8%. Es un indicador positivo, pero indica la necesidad de mejorar la experiencia de usuario.

El ejemplo que hemos desarrollado ha sido para una variable cuantitativa discreta, pero podría replicarse para cualquier variable cualitativa, incluso entre pares de variables cualitativas. Un ejemplo: queremos estudiar el número de pólizas vendidas para cada región geográfica según los tramos de edad.

	Norte y noroeste	Centro	Sur	Levante	Noreste	Islas
18-30 años	31	80	45	63	43	28
31-60 años	149	198	166	143	175	110
>60 años	159	200	180	130	179	115

Para calcular el resto de métricas se siguen los mismos pasos que en el caso anterior, pero en este caso cada métrica tendrá una tabla diferente. Este tipo de tablas, a menudo, aparecen representadas como [heatmaps](#).

Outliers

X Edix Educación

Dentro de los posibles valores que toman las variables, en ocasiones, hay **valores atípicos o outliers**, es decir, observaciones (o individuos) que distan del resto de los datos. Los outliers no tienen que ser exclusivamente ‘cuantitativos’, pueden ser también ‘cualitativos’. Veamos algunos ejemplos:

- Los **GRPS semanales** de una marca oscilan entre 50–150GRPs, pero hay una semana que tiene un valor de 600 GRPs. Este dato probablemente sea un outlier, por lo que convendría confirmar con la agencia de medios o el departamento de marketing.
- Estamos estudiando la **tasa de paro de los países europeos** con valores entre el 15% y el 40%. Uno de los países muestra un valor de -3%. Claramente es un error ya que es imposible que la tasa de paro sea negativa.
- Estamos estudiando las **plataformas** en las que hemos hecho presión publicitaria, y encontramos una web con una única impresión. En este caso, esa observación se eliminaría del estudio, previa confirmación con el departamento digital.

Los valores atípicos son aquellos que si los eliminamos la distribución cambia, y por tanto, distorsionan la imagen de la población que están representando (recordemos la sensibilidad a outliers de la media, mediana, varianza y desviación típica).

Para tratar los outliers, primero hay que **detectar si son valores atípicos o errores**. Una vez se conoce el origen, en función del estudio, importancia de la observación... se decidirá eliminar la observación o imputar al valor adecuado.

Los gráficos son posiblemente la forma más rápida y eficaz de detectar outliers. Los boxplots son los gráficos por excelencia para detectar outliers, aunque también es muy frecuente usar gráficos de dispersión, de líneas o histogramas.

En el caso de contar con datos que tengan fecha, lo más recomendable es graficarlos como si fuera una serie temporal (es decir, la fecha en el eje X, la variable que queramos entender en el eje Y, y la serie unida con una línea).

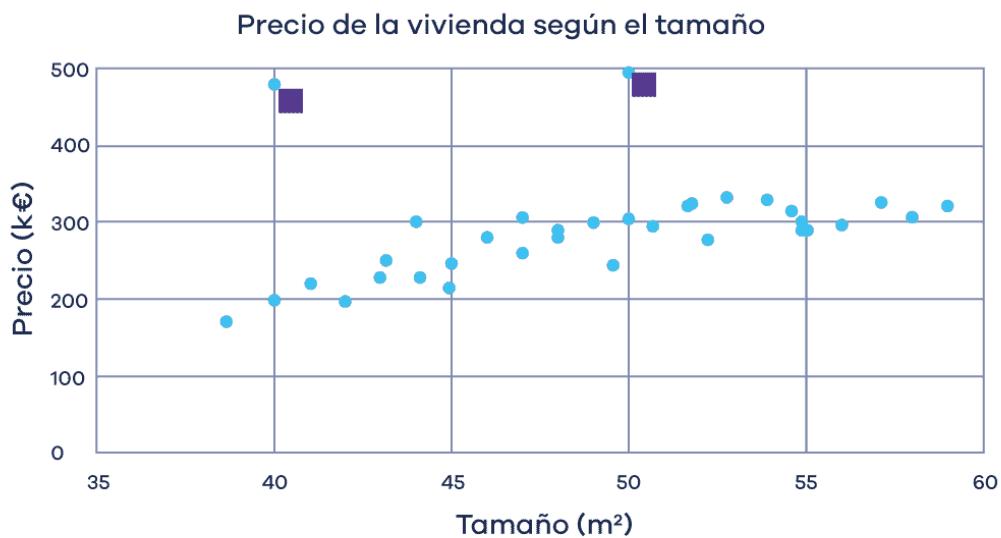
En general, tanto para detectar outliers como para conocer la variabilidad de las variables es muy recomendable graficar la serie de publicidad que nos interesa junto con la variable objetivo que queremos estudiar. Esto nos va a ayudar a detectar de forma más clara si es un outlier o no.

Veamos un ejemplo.

Queremos analizar la relación entre el precio de venta y el tamaño de las viviendas

De forma rápida, observamos que las dos viviendas con un cuadrado violeta probablemente sean outliers, ya que sus características son diferentes a las de ese mismo tamaño. Primero se revisará la ficha de la vivienda de estas dos observaciones y, en caso de no haber errores en los datos, se eliminarán esos registros, ya que la información de estas dos viviendas es errónea y podría conducirnos a conclusiones equivocadas.

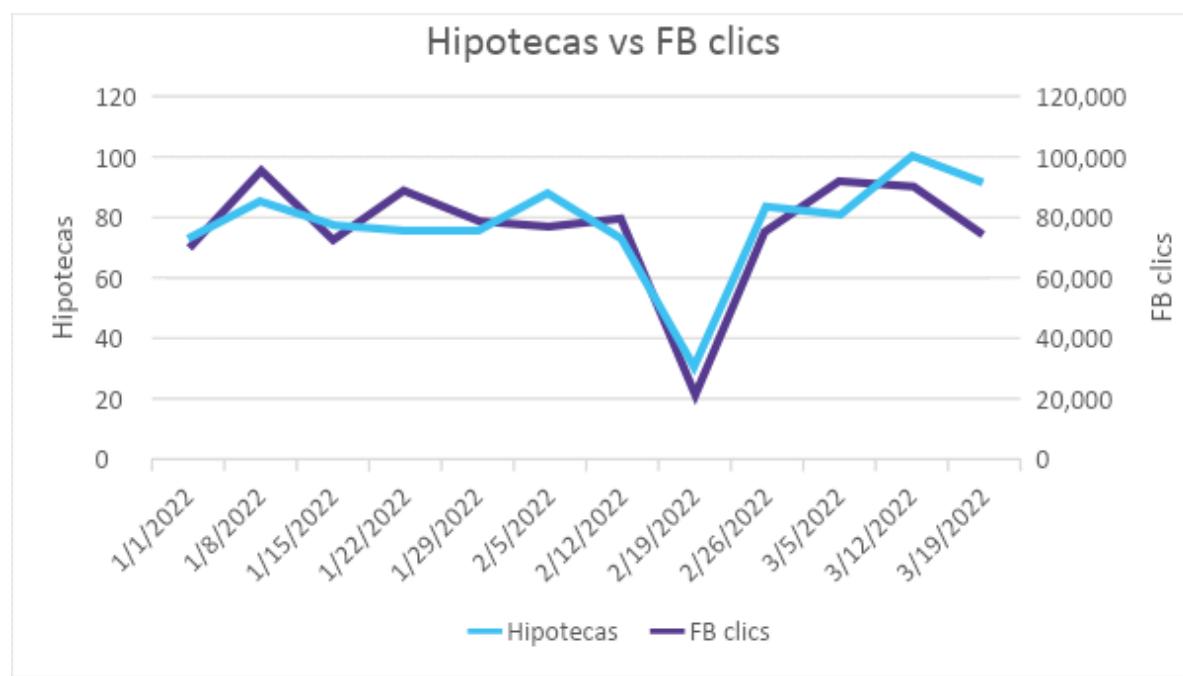
Si hacemos un estudio con estas dos observaciones incluidas, el precio medio de la vivienda quedará muy sesgado y los resultados no serán realistas. Como ves, hemos analizado las dos series para entender si había outliers. Con solo la de tamaño probablemente no los hubiéramos distinguido.



Lo comprenderemos mejor con otro ejemplo

En este caso, queremos estudiar los outliers de FB clics. Podríamos haber graficado el evolutivo de la serie solamente de FB clics, pero la enfrentamos a la variable que queremos estudiar, que aquí son las hipotecas.

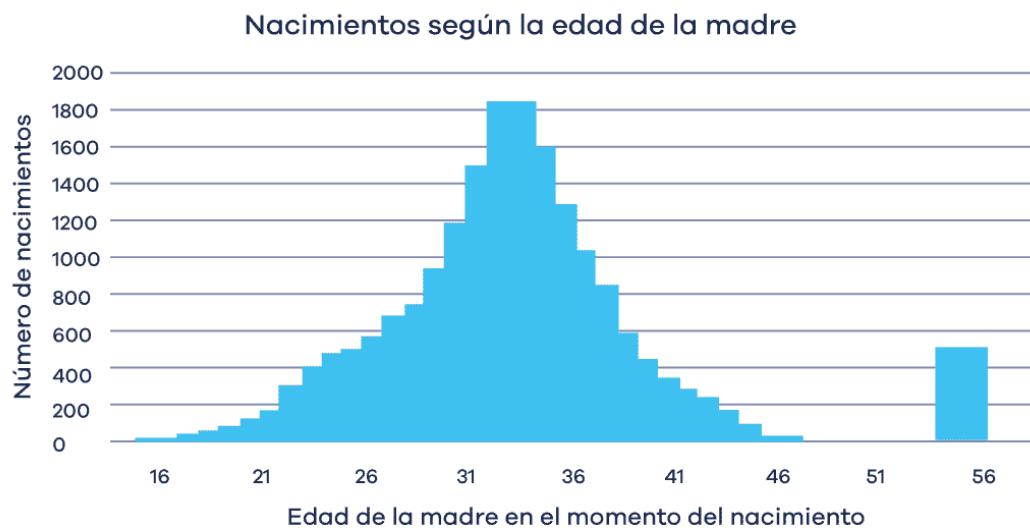
En el gráfico, el 19/02/2022 se puede ver claramente un pico. Como tenemos las dos series, podemos decir que no es un outlier. En caso de que no hubiéramos incluido la variable hipotecas, probablemente habríamos detectado en ambas series este punto como un outlier.



Otro ejemplo: supongamos que queremos estudiar el número de nacimientos según la edad de la madre.

Calculamos la tabla de frecuencias, y graficamos el resultado en un histograma.

Se ve claramente un valor atípico en la edad de 55 años. Esto puede ser debido a que la edad máxima que permite almacenar el sistema sea 55 y ante el desconocimiento se impute automáticamente a este valor. Para trabajar con esta serie se cortará el histórico para quedarnos con mujeres de 16 a 46 años, de tal forma que los valores 55 no afecten a nuestro estudio.



Además de la forma gráfica, existen múltiples métodos para la detección de estas observaciones anómalas: criterio de Chauvenet, criterio de Pierce, distancia de Cook, rango intercuartílico, basado en la desviación típica, basado en la mediana de las desviaciones absolutas (MEDA), etc. Además del soporte gráfico, otro buen soporte es el rango intercuartílico.

Rango intercuartílico

El **método de Tukey**, comúnmente conocido como el **rango intercuartílico**, es la **técnica matemática para detectar outliers más utilizada y fácil de calcular**. Como aprendimos en el fastbook anterior con las medidas de posición, está basada en la diferencia entre Q_3 y Q_1 . Diremos que un valor q es atípico si

$$q < Q_1 - 1,5 * IQR \text{ o } q > Q_3 + 1,5 * IQR$$

En ocasiones, el valor 1,5 se sustituye por 3 si queremos solamente eliminar los outliers externos, o si la distribución tiene colas muy largas.

Continuando con el ejemplo de las medidas de posición, recordamos que tenemos las siguientes calificaciones:

Asignatura	id1	id2	id3	id4	id5	id6	id7	id8	id9	id10
Calificación	6	9	8	0	5	4	7	10	6	7

Como ya calculamos, los cuartiles son $Q_1=5$, $Q_2=6,5$, y $Q_3=8$. El rango intercuartílico, por tanto, sería $IQR=Q_3-Q_1=8-5=3$.

Los límites para considerar un valor como outlier son:

$$Q_1 - 1,5 * IQR = 5 - 1,5 * 3 = 5 - 4,5 = 0,5$$

$$Q_3 + 1,5 * IQR = 8 + 1,5 * 3 = 8 + 4,5 = 12,5$$

En este caso, solamente la calificación de la asignatura id4 con 0 será un valor outlier, confirmando que esta nota no está alineada con el resto ya que el alumno no se presentó al examen.

Covarianza

X Edix Educación

La covarianza es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias. Nos permite saber cómo se comporta una variable X en función de lo que hace otra variable Y. Es decir, cuando X sube, ¿cómo se comporta Y?

La covarianza se calcula como:

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{1}{n-1} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}$$

La covarianza(X,Y) puede tomar valores positivos y negativos, de tal forma que:

- Si $\text{cov}(X,Y) < 0$ hay una relación negativa (cuando X sube, Y baja).
- Si $\text{cov}(X,Y) > 0$ hay una relación positiva (cuando X sube, Y sube).

Por ejemplo: supongamos que queremos estudiar el grado de correlación de los 12 alumnos de una clase que tienen estadística y matemáticas. Las calificaciones obtenidas para cada alumno son:

Alumno	Estadística (X)	Matemáticas (Y)
al1	2	1
al2	3	3
al3	4	2
al4	4	4
al5	5	4
al6	6	4
al7	6	6
al8	7	4
al9	7	6
al10	8	7
al11	10	9
al12	10	10

Añadimos una nueva columna que sea $x_i * y_i$

Alumno	Estadística (X)	Matemáticas (Y)	$x_i * y_i$
al1	2	1	2
al2	3	3	9
al3	4	2	8
al4	4	4	16
al5	5	4	20
al6	6	4	24
al7	6	6	36
al8	7	4	28
al9	7	6	42
al10	8	7	56
al11	10	9	90
al12	10	10	100
Total (suma)	72	60	431

Calculamos X=6 e Y=5.

De esta forma:

$$cov(X,Y) = \frac{1}{n-1} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{1}{12-1} 431 - 6 * 5 = 9.18$$

Al ser **positiva**, nos indica que a notas altas en estadística también habrá notas altas en matemáticas.

Outliers

X Edix Educación

Dentro de los posibles valores que toman las variables, en ocasiones, hay **valores atípicos o outliers**, es decir, observaciones (o individuos) que distan del resto de los datos. Los outliers no tienen que ser exclusivamente ‘cuantitativos’, pueden ser también ‘cualitativos’. Veamos algunos ejemplos:

- Los **GRPS semanales** de una marca oscilan entre 50–150GRPs, pero hay una semana que tiene un valor de 600 GRPs. Este dato probablemente sea un outlier, por lo que convendría confirmar con la agencia de medios o el departamento de marketing.
- Estamos estudiando la **tasa de paro de los países europeos** con valores entre el 15% y el 40%. Uno de los países muestra un valor de -3%. Claramente es un error ya que es imposible que la tasa de paro sea negativa.
- Estamos estudiando las **plataformas** en las que hemos hecho presión publicitaria, y encontramos una web con una única impresión. En este caso, esa observación se eliminaría del estudio, previa confirmación con el departamento digital.

Los valores atípicos son aquellos que si los eliminamos la distribución cambia, y por tanto, distorsionan la imagen de la población que están representando (recordemos la sensibilidad a outliers de la media, mediana, varianza y desviación típica).

Para tratar los outliers, primero hay que **detectar si son valores atípicos o errores**. Una vez se conoce el origen, en función del estudio, importancia de la observación... se decidirá eliminar la observación o imputar al valor adecuado.

Los gráficos son posiblemente la forma más rápida y eficaz de detectar outliers. Los boxplots son los gráficos por excelencia para detectar outliers, aunque también es muy frecuente usar gráficos de dispersión, de líneas o histogramas.

En el caso de contar con datos que tengan fecha, lo más recomendable es graficarlos como si fuera una serie temporal (es decir, la fecha en el eje X, la variable que queramos entender en el eje Y, y la serie unida con una línea).

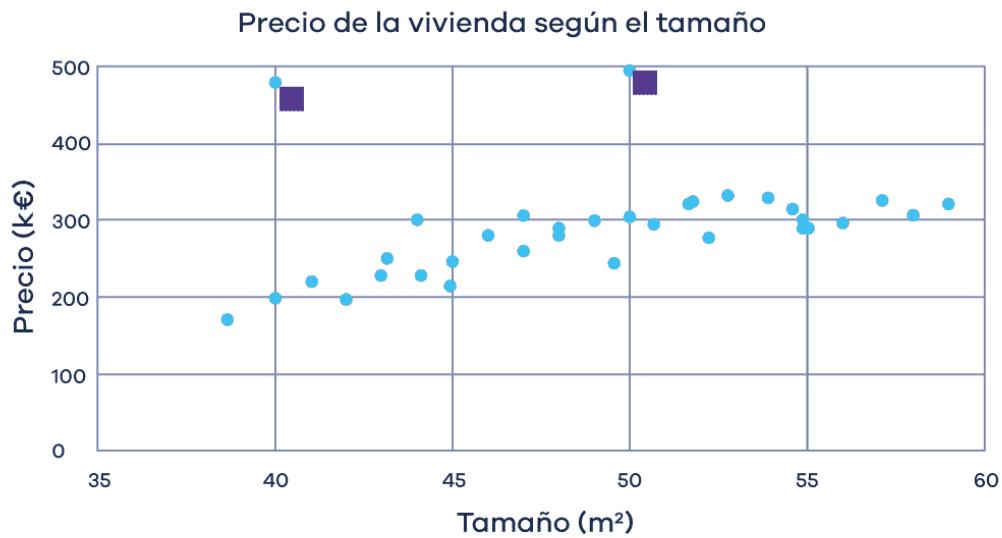
En general, tanto para detectar outliers como para conocer la variabilidad de las variables es muy recomendable graficar la serie de publicidad que nos interesa junto con la variable objetivo que queremos estudiar. Esto nos va a ayudar a detectar de forma más clara si es un outlier o no.

Veamos un ejemplo.

Queremos analizar la relación entre el precio de venta y el tamaño de las viviendas

De forma rápida, observamos que las dos viviendas con un cuadrado violeta probablemente sean outliers, ya que sus características son diferentes a las de ese mismo tamaño. Primero se revisará la ficha de la vivienda de estas dos observaciones y, en caso de no haber errores en los datos, se eliminarán esos registros, ya que la información de estas dos viviendas es errónea y podría conducirnos a conclusiones equivocadas.

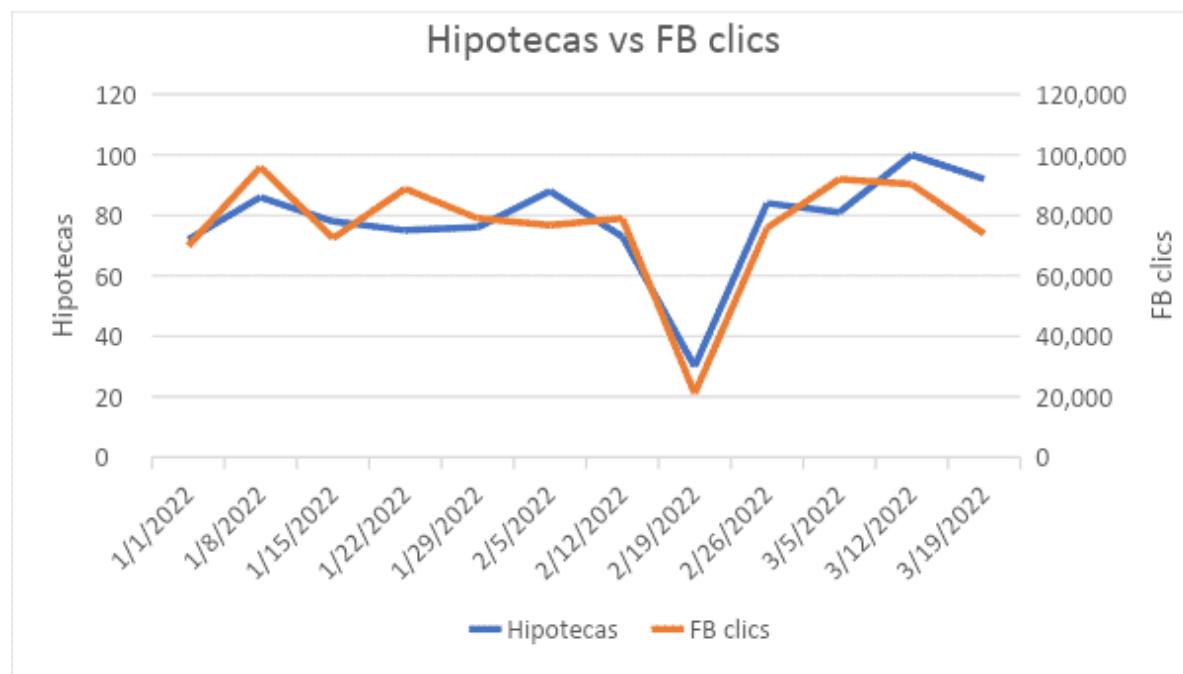
Si hacemos un estudio con estas dos observaciones incluidas, el precio medio de la vivienda quedará muy sesgado y los resultados no serán realistas. Como ves, hemos analizado las dos series para entender si había outliers. Con solo la de tamaño probablemente no los hubiéramos distinguido.



Lo comprenderemos mejor con otro ejemplo

En este caso, queremos estudiar los outliers de FB clics. Podríamos haber graficado el evolutivo de la serie solamente de FB clics, pero la enfrentamos a la variable que queremos estudiar, que aquí son las hipotecas.

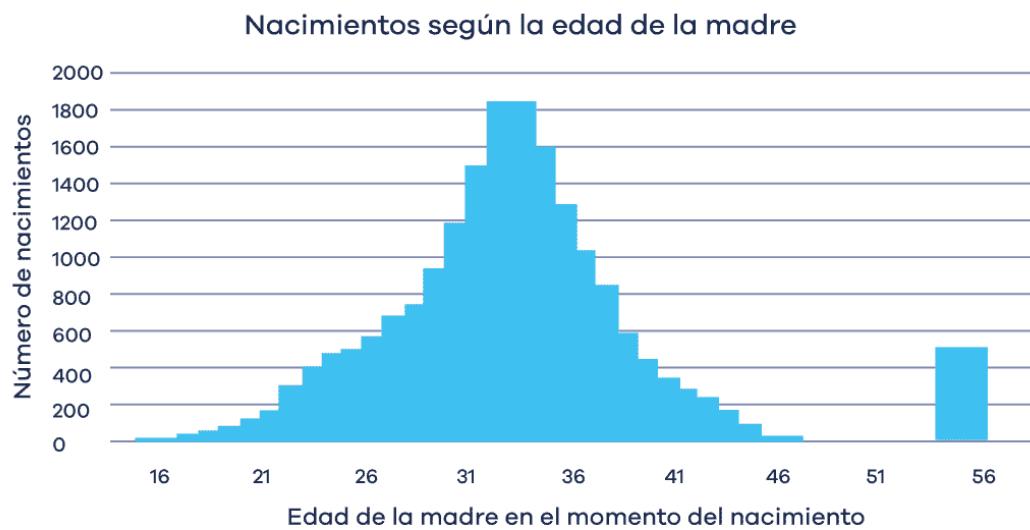
En el gráfico, el 19/02/2022 se puede ver claramente un pico. Como tenemos las dos series, podemos decir que no es un outlier. En caso de que no hubiéramos incluido la variable hipotecas, probablemente habríamos detectado en ambas series este punto como un outlier.



Otro ejemplo: supongamos que queremos estudiar el número de nacimientos según la edad de la madre.

Calculamos la tabla de frecuencias, y graficamos el resultado en un histograma.

Se ve claramente un valor atípico en la edad de 55 años. Esto puede ser debido a que la edad máxima que permite almacenar el sistema sea 55 y ante el desconocimiento se impute automáticamente a este valor. Para trabajar con esta serie se cortará el histórico para quedarnos con mujeres de 16 a 46 años, de tal forma que los valores 55 no afecten a nuestro estudio.



Además de la forma gráfica, existen múltiples métodos para la detección de estas observaciones anómalas: criterio de Chauvenet, criterio de Pierce, distancia de Cook, rango intercuartílico, basado en la desviación típica, basado en la mediana de las desviaciones absolutas (MEDA), etc. Además del soporte gráfico, otro buen soporte es el rango intercuartílico.

Rango intercuartílico

El **método de Tukey**, comúnmente conocido como el **rango intercuartílico**, es la **técnica matemática para detectar outliers más utilizada y fácil de calcular**. Como aprendimos en el fastbook anterior con las medidas de posición, está basada en la diferencia entre Q_3 y Q_1 . Diremos que un valor q es atípico si

$$q < Q_1 - 1,5 \cdot IQR \text{ o } q > Q_3 + 1,5 \cdot IQR$$

En ocasiones, el valor 1,5 se sustituye por 3 si queremos solamente eliminar los outliers externos, o si la distribución tiene colas muy largas.

Continuando con el ejemplo de las medidas de posición, recordamos que tenemos las siguientes calificaciones:

Asignatura	id1	id2	id3	id4	id5	id6	id7	id8	id9	id10
Calificación	6	9	8	0	5	4	7	10	6	7

Como ya calculamos, los cuartiles son $Q_1=5$, $Q_2=6,5$, y $Q_3=8$. El rango intercuartílico, por tanto, sería $IQR=Q_3-Q_1=8-5=3$.

Los límites para considerar un valor como outlier son:

$$Q_1 - 1,5 \cdot IQR = 5 - 1,5 \cdot 3 = 5 - 4,5 = 0,5$$

$$Q_3 + 1,5 \cdot IQR = 8 + 1,5 \cdot 3 = 8 + 4,5 = 12,5$$

En este caso, solamente la calificación de la asignatura id4 con 0 será un valor outlier, confirmando que esta nota no está alineada con el resto ya que el alumno no se presentó al examen.

Covarianza

X Edix Educación

La covarianza es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias. Nos permite saber cómo se comporta una variable X en función de lo que hace otra variable Y. Es decir, cuando X sube, ¿cómo se comporta Y?

La covarianza se calcula como:

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{1}{n-1} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}$$

La covarianza(X,Y) puede tomar valores positivos y negativos, de tal forma que:

- Si $\text{cov}(X,Y) < 0$ hay una relación negativa (cuando X sube, Y baja).
- Si $\text{cov}(X,Y) > 0$ hay una relación positiva (cuando X sube, Y sube).

Por ejemplo: supongamos que queremos estudiar el grado de correlación de los 12 alumnos de una clase que tienen estadística y matemáticas. Las calificaciones obtenidas para cada alumno son:

Alumno	Estadística (X)	Matemáticas (Y)
al1	2	1
al2	3	3
al3	4	2
al4	4	4
al5	5	4
al6	6	4
al7	6	6
al8	7	4
al9	7	6
al10	8	7
al11	10	9
al12	10	10

Añadimos una nueva columna que sea $x_i * y_i$

Alumno	Estadística (X)	Matemáticas (Y)	$x_i * y_i$
al1	2	1	2
al2	3	3	9
al3	4	2	8
al4	4	4	16
al5	5	4	20
al6	6	4	24
al7	6	6	36
al8	7	4	28
al9	7	6	42
al10	8	7	56
al11	10	9	90
al12	10	10	100
Total (suma)	72	60	431

Calculamos X=6 e Y=5.

De esta forma:

$$cov(X,Y) = \frac{1}{n-1} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{1}{12-1} 431 - 6 * 5 = 9.18$$

Al ser **positiva**, nos indica que a notas altas en estadística también habrá notas altas en matemáticas.

Correlación

X Edix Educación

La correlación es la medida por excelencia entre dos variables numéricas para determinar la relación entre ellas.

El análisis de correlación mide la fuerza y dirección entre dos variables cuantitativas continuas.

Decimos que dos variables están correlacionadas cuando **comparten comportamientos de tendencia** de manera directa o inversa, es decir, mide la dependencia de una variable con respecto de otra. La correlación se puede medir con el **coeficiente de Pearson** y el **coeficiente de Spearman**.

Recordemos que una de las formas más extendidas para definir el nivel más adecuado de ad-stock es con el uso de la correlación. Verás cómo ahora tienes mucho más claro qué nivel debes elegir una vez conozcas esta medida.

1

Coeficiente de correlación de Pearson

El coeficiente de **correlación de Pearson** es una medida de **dependencia lineal** entre dos variables aleatorias cuantitativas continuas, independiente de la escala de medida de las variables. Mide el grado de variación lineal conjunta.

Si se representan en un diagrama de dispersión los valores que toman dos variables, el coeficiente de correlación lineal señalará lo bien o lo mal que el conjunto de puntos representados se aproxima a una recta.

Dando X, Y dos variables aleatorias sobre una población, el coeficiente de correlación de Pearson es:

$$r_{xy} = \frac{\sigma_{XY}}{\sigma_x \sigma_Y} = \frac{cov(X,Y)}{\sqrt{var(X) * var(Y)}}$$

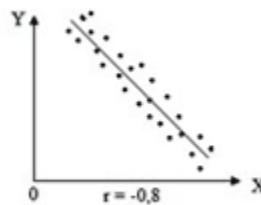
Donde:

- σ_{XY} es la covarianza de (X,Y).
- σ_X es la desviación estándar de la variable X.
- σ_Y es la desviación estándar de la variable Y.

Igualmente, podemos calcular el coeficiente de correlación de Pearson sobre un espacio muestral de forma que:

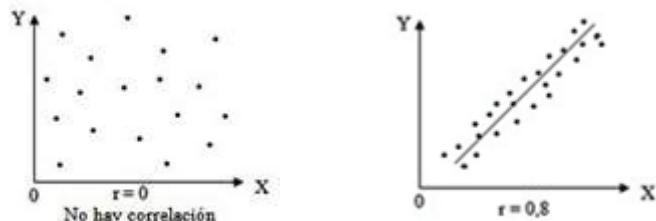
$$r_{xy} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{(n-1)s_x s_y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}.$$

El coeficiente de correlación de Pearson toma valores entre **-1 y 1**:



Si toma valores cercanos a **-1**, la correlación es fuerte e inversa.

Si toma valores cercanos a **0**, la correlación es débil.



Si toma valores cercanos a **1**, la correlación es fuerte y directa.

Diremos que la relación es directa (correlación positiva) si cuando una aumenta, la otra también lo hace; o inversa (correlación negativa) si cuando una aumenta, la otra disminuye.

Si continuamos con el ejemplo anterior partiendo de la tabla con $x_i * y_i$, calculamos x_i^2 e y_i^2 :

Alumno	Estadística (X)	Matemáticas (Y)	$x_i * y_i$	x_i^2	y_i^2
al1	2	1	2	4	1
al2	3	3	9	9	9
al3	4	2	8	16	4
al4	4	4	16	16	16
al5	5	4	20	25	16
al6	6	4	24	36	16
al7	6	6	36	36	36
al8	7	4	28	49	16
al9	7	6	42	49	36
al10	8	7	56	64	49
al11	10	9	90	100	81
al12	10	10	100	100	100
Total (suma)	72	60	431	504	380

Sustituyendo en la fórmula obtendríamos:

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} * \sqrt{n \sum y_i^2 - (\sum y_i)^2}} = \frac{12*431-72*60}{\sqrt{12*504-72^2} * \sqrt{12*380-60^2}} = 0,9355$$

El coeficiente de correlación es muy alto, por lo que podemos concluir que ambas variables están muy relacionadas.

Las principales herramientas informáticas con las que trabajarás en un futuro tienen su cálculo implementado mediante un comando parecido a COR().

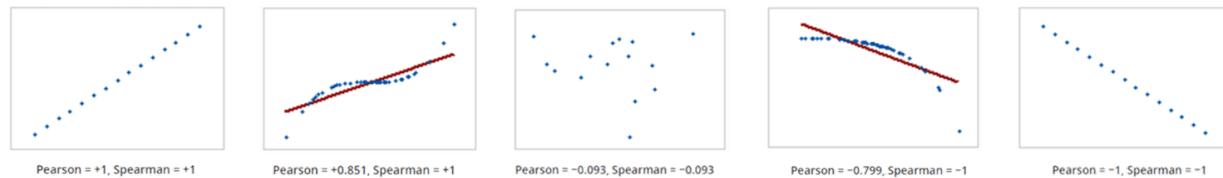
2

Coefficiente de correlación de Spearman

A diferencia de la relación lineal de la correlación de Pearson, el coeficiente de **correlación de Spearman** se centra en la **relación monótona**, es decir, las variables cambian en tiempo y dirección (directa o inversa), pero no necesariamente a un ritmo constante. Los posibles valores y la interpretación del coeficiente es la misma que para Pearson.

En este caso, la fórmula matemática es más complicada, por lo que con saber el significado y cómo calcularlo en el programa informático que uses es suficiente (suele llamarse mediante COR() y dentro se especifica el método de cálculo).

Con el siguiente gráfico podrás identificar diferentes situaciones en los que los coeficientes de correlación de Pearson y Spearman pueden comportarse igual o diferente.



3

Matriz de correlación

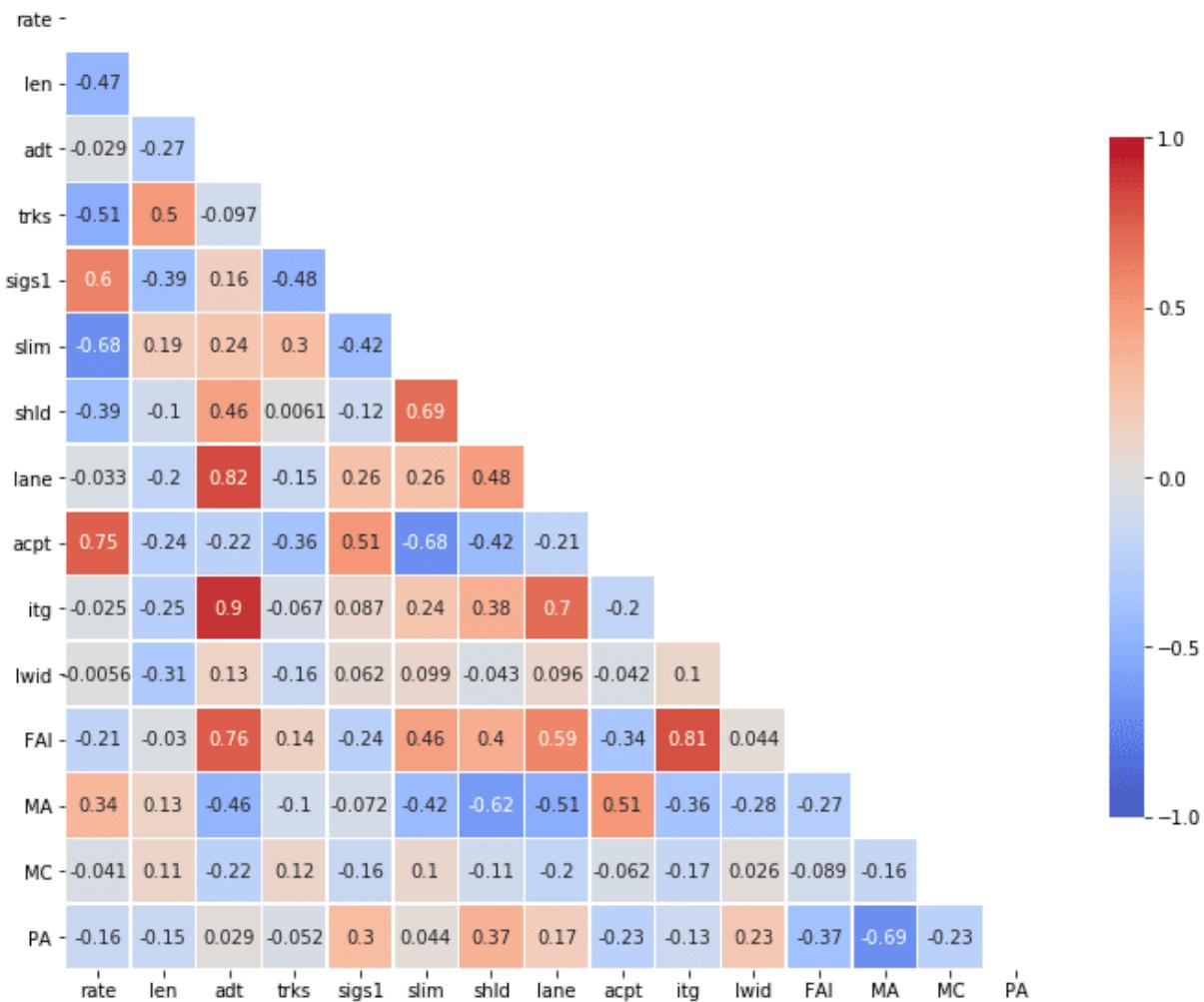
Cuando queremos calcular múltiples correlaciones entre pares de variables se suele apoyar en la matriz de correlaciones.

Una matriz de correlaciones es un modo de resumir y localizar correlaciones entre **múltiples variables**.

Algunas de las **características** que presenta son:

- Es **simétrica respecto a la diagonal**.
- La **diagonal** son todo **1**, ya que una variable consigo misma tiene correlación 1.
- En cada elemento (k, i) de la matriz, tenemos la **correlación entre las variables** que corresponden a la fila k y la columna i .
- Se suele representar como **triangular inferior o superior** como si fuera un heatmap para que de un primer vistazo identificar los valores más altos.

Un ejemplo de una matriz de correlación triangular inferior es:



Análisis exploratorio de datos (EDA)

X Edix Educación

El análisis exploratorio de datos es el conjunto de técnicas que se aplican a un conjunto de datos para entender el comportamiento de cada una de las variables de forma individual y de las interacciones entre ellas.

Un buen análisis de datos es muy importante para una correcta compresión de la distribución de los datos.

La **combinación de las medidas**, que aprendimos con el fastbook anterior, junto con las técnicas de este, nos proporcionan las herramientas suficientes para hacer un buen análisis exploratorio (EDA, *exploratory data analysis*). Para llevarlo a cabo, nos apoyaremos en algunos gráficos, ya que son recursos imprescindibles para presentar la información y entender y analizar los datos de forma fácil y rápida.

Realizaremos un EDA de forma práctica desde la recepción de los datos hasta las conclusiones (el ejemplo está basado en el que vimos de las tablas de frecuencia, pero con la base de datos enriquecida).

1

Recogida de datos

Supongamos que queremos medir la lealtad de 50 de nuestros clientes en relación a un producto. Además, queremos ver la relación que tiene la edad y el carrito medio de compra (gasto promedio de compra) de cada cliente sobre la valoración.

Los datos que tenemos son:

Individuo	id1	id2	id3	id4	id5	id6	id7	id8	id9	id10	id11	id12	id13
Valoración	9	10	5	7	8	7	6	7	8	0	8	0	9
Edad	34	68	75	43	53	55	70	38	42	48	47	69	57
Compra promedio	8€	55€	74€	43€	37€	57€	590€	34€	39€	25€	41€	40€	13€

Individuo	id14	id15	id16	id17	id18	id19	id20	id21	id22	id23	id24	id25	id26
Valoración	10	9	8	5	8	8	9	7	8	7	9	8	9
Edad	50	27	31	65	24	45	60	35	56	39	29	28	31
Compra promedio	50€	5€	37€	18€	13€	44€	25€	33€	26€	45€	37€	48€	15€

Individuo	id27	id28	id29	id30	id31	id32	id33	id34	id35	id36	id37	id38	id39
Valoración	7	4	9	9	7	4	10	6	8	9	10	9	10
Edad	50	32	31	57	64	19	44	62	61	2	41	18	34
Compra promedio	18€	20€	9€	39€	46€	5€	51€	53€	33€	59€	33€	20€	7€

Individuo	id40	id41	id42	id43	id44	id45	id46	id47	id48	id49	id50
Valoración	2	6	7	10	3	9	8	5	7	8	5
Edad	52	25	30	33	44	47	43	26	79	81	58
Compra promedio	30€	44€	60€	16€	46€	29€	10€	30€	45€	35€	37€

2

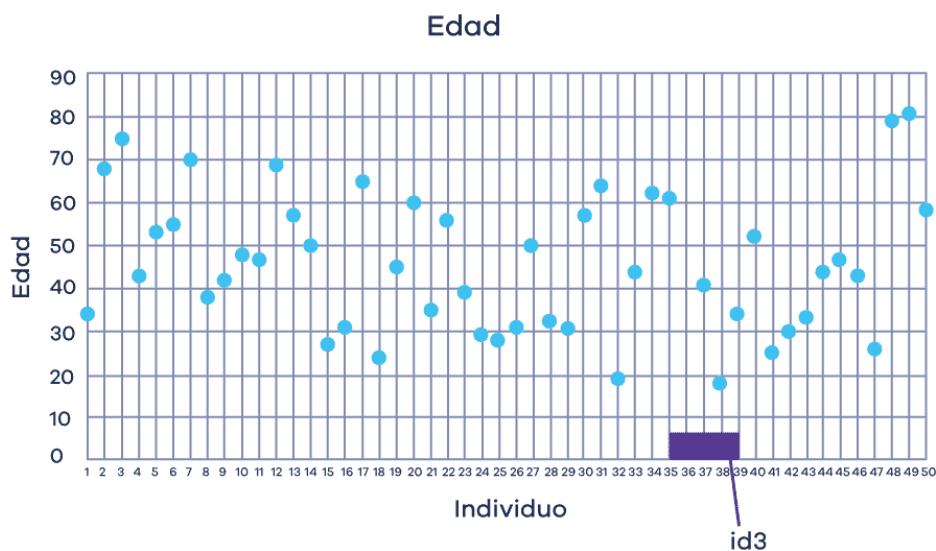
Validación de datos

En el primer paso debemos confirmar que los datos son correctos, por lo que para ello estudiamos su distribución individual y **detectamos outliers**.

Empezaremos estudiando la **variable objetivo: valoración**. Al ser una variable categórica para comprobar que los valores que toma son correctos, solamente hace falta quedarse con los valores únicos y chequear que son de 0-10 (en base de datos el comando suele ser *unique*).

Lo hacemos y verificamos que la variable no tiene atípicos.

A continuación, estudiamos el **comportamiento de la variable edad**. La graficaremos mediante un **gráfico de dispersión**, en el que el eje X sea cada individuo y en el eje Y tengamos la edad.

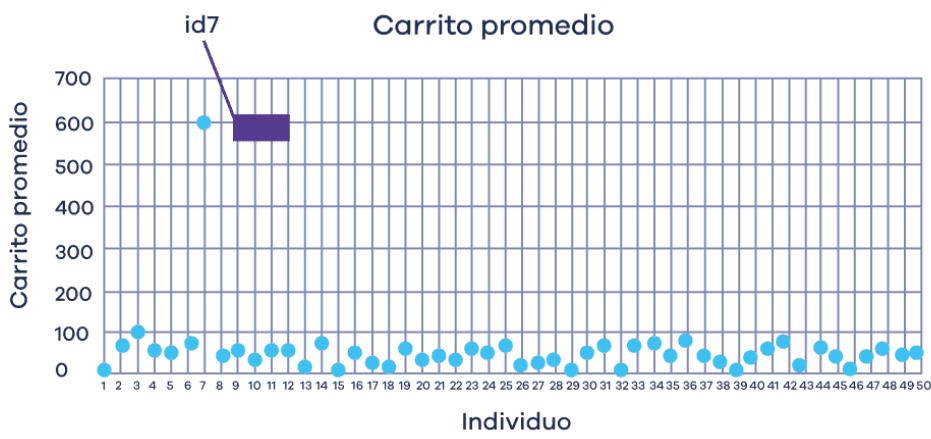


Como podemos apreciar, hay una observación (id36) que puede presentar un outlier, ya que la edad es esta observación: de 2 años de edad. Chequeamos mediante el rango intercuartílico:

$$q=2>Q_1-1,5IQR=31,25-1,5*57-31,25=-7,3$$

A pesar de que la observación id36 no es un outlier, una persona de 2 años no ha podido hacer la encuesta por lo que eliminamos la observación.

Por último, estudiamos la distribución de **carrito medio**. Seguimos los mismos pasos que para edad.



Como podemos ver, el individuo id7 que tiene un carrito medio de 590€, mucho más alto que el resto. Chequeamos mediante el rango intercuartílico que se trata de un outlier:

$$q=590\text{€}>Q_3+1,5IQR=45+1,5*45-20=57,3\text{€}$$

Como podíamos suponer, el individuo tiene un carrito promedio mucho más alto que el límite superior, por lo que confirmamos que es un outlier y eliminamos la observación.

3

Comportamiento individual e interacción de las variables

Una vez que tenemos la base de datos limpia, pasamos a estudiar el comportamiento de la valoración, edad y carrito promedio frente a la valoración, y de las interacciones entre ambas.

Empezaremos estudiando variable objetivo: valoración. Primero crearemos la tabla de frecuencia con todas las métricas, y luego nos apoyaremos en los histogramas —representan la frecuencia absoluta (N_i)—, y gráficos de tarta —frecuencia relativa en porcentaje (f_i)—.

Tabla de frecuencia

Creamos la tabla de frecuencia correspondiente según los pasos ya aprendidos. A la vista de los resultados, destacar que la moda (y la mediana) es 8 con un 23% de las respuestas (f_i). En paralelo, los resultados parecen positivos al tener tan solo un 25% de detractores (F_i).

Calculando el indicador NPS, confirmamos que es de $34 - 25 = 9$ y por tanto positiva.

Puntuación	Frecuencia	Frecuencia absoluta acumulada (N_i)	Frecuencia relativa en porcentaje (f_i)	Frecuencia relativa acumulada (F_i)
0	2	2	4%	4%
1	0	2	0%	4%
2	1	3	2%	6%
3	1	4	2%	8%
4	2	6	4%	13%
5	4	10	8%	21%
6	2	12	4%	25%
7	9	21	19%	44%
8	11	32	23%	67%
9	10	42	21%	88%
10	6	48	13%	100%
Total	48		100%	

Histograma

Visualizamos los datos primero con un **histograma** (gráficos de barras donde la longitud de cada barra representa la frecuencia de aparición de los valores).

A la vista de los resultados, se puede ver claramente que los encuestados se condensan mayoritariamente entre los valores 7, 8 y 9. En contraposición, destaca la baja frecuencia de valores de 1 a 3.

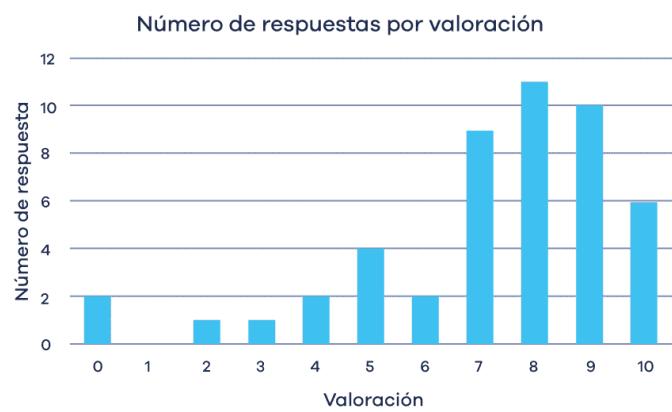
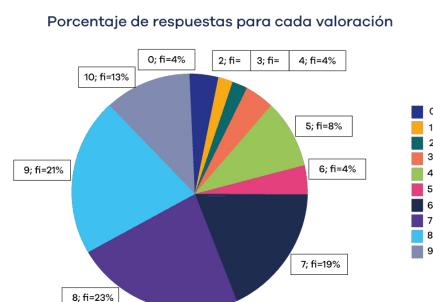


Gráfico de tarta

Los **gráficos de tarta** son gráficos circulares muy útiles para representar porcentajes o proporciones. Cada ‘trozo’ de la tarta representa la frecuencia relativa de cada valor.

De un simple vistazo, se puede ver que la tarta tiene mayoritariamente colores verdes, y que la suma de las puntuaciones 8, 9 y 10 representan más del 50% del total. Las puntuaciones de 6 o inferior representan aproximadamente el 25% del total.

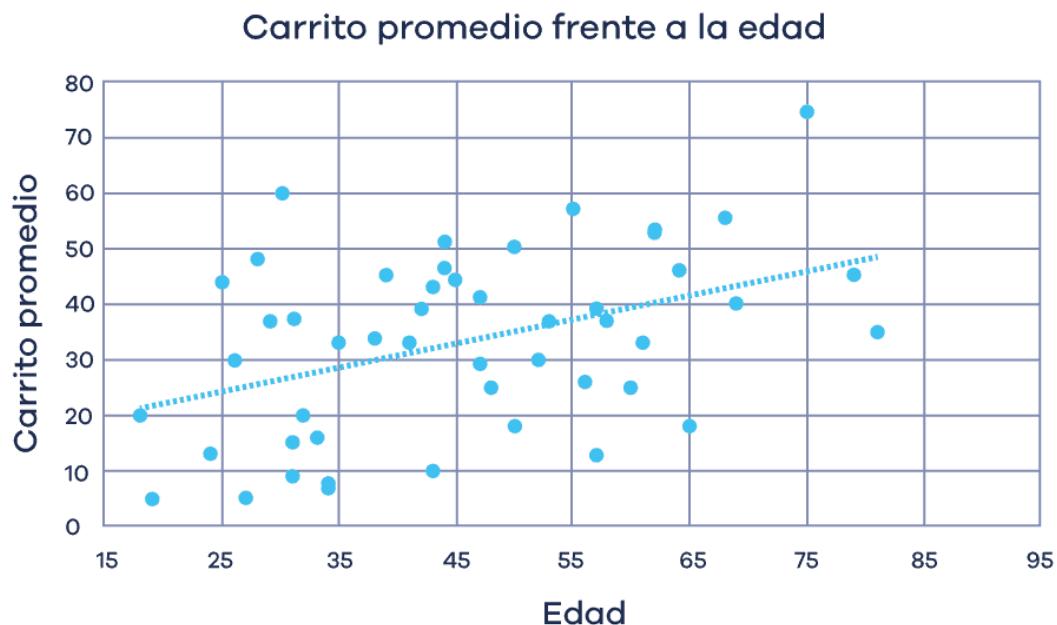


Pasemos ahora a estudiar el **comportamiento de edad y el carrito promedio frente a la valoración y entre ellas**. Para no quedarnos sin observaciones, agruparemos las valoraciones en detractores, pasivos y promotores.

Primero calculamos la relación entre **edad y carrito promedio**. Para ello, graficamos mediante un gráfico de dispersión, y calculamos el coeficiente de correlación.

La correlación entre ambas es de 0,42.

A la vista de los resultados, observamos con el gráfico que hay una relación positiva creciente entre ambas, pero con mucha dispersión alrededor de la recta. Esto lo confirmamos con el coeficiente de correlación que es bajo: 0,42.



Empezamos por estudiar la relación entre la **valoración** (agrupada) y la **edad**. Calculamos la media, mediana y desviación estándar de la edad para toda la muestra y de forma particular para cada grupo.

A la vista de los resultados vemos que el grupo de detractores son más jóvenes que el resto de grupos y que la media y mediana general. De igual forma, este grupo presenta una menor dispersión de los datos alrededor de la media

	Media	Mediana	SD
Promotores	47,9	50,0	18,8
Pasivos	47,2	44,0	15,6
Detractores	41,3	37,5	14,0
Detractores	45,4	44,0	15,9

Replicamos los mismos cálculos para el **carrito promedio frente a la valoración** (agrupada).

Observamos que el carrito promedio de los detractores es considerablemente más bajo en media y mediana que el resto de grupos y del total, aunque presenta una dispersión más alta del promedio. Este comportamiento es muy interesante, puesto que los clientes que gastan menos son los que peor consideración tienen de nuestra marca.

	Media	Mediana	SD
Promotores	35,2	33,5	18,2
Pasivos	37,2	38,0	13,0
Detractores	25,8	22,5	16,8
Total	32,9	34,5	16,2

4

Conclusiones

Como **resumen del análisis exploratorio**, podemos decir que a nivel general la lealtad de los clientes es buena, con un indicador NPS 7 con clientes muy diversos en edad y en carrito promedio.

Evaluando cada grupo, observamos que los promotores y los pasivos tienen una distribución similar de edad y carrito promedio, aunque destaca la dispersión de los promotores en ambas variables.

Por otro lado, el grupo de los detractores es el más diferente al resto, donde se agrupan los compradores más jóvenes y con carritos más baratos y más centrados en la media que el resto.

Se recomienda llevar a cabo **medidas para mejorar la satisfacción de los clientes**, que están centrados en perfiles más jóvenes y con compras de menor importe.

Resumen

X Edix Educación

Hagamos un repaso de los conceptos dados en este fastbook:

Tablas de frecuencia

Hemos aprendido la forma de calcular las **tablas de frecuencia** y la utilidad que tienen. Sabemos calcular las diferentes métricas asociadas y como graficarlas para poder extraer conclusiones.

Outlier

Nos hemos familiarizado con el término de **outlier (o atípico)**, su significado y el impacto que tiene sobre un conjunto de datos. Hemos conocido diversas formas de detectarlo mediante gráficos y técnicas.

Covarianza

Hemos descubierto el significado de **covarianza** y la forma de calcularlo.

Correlación

Hemos aprendido a medir la relación entre las variables mediante el coeficiente de correlación. Sabemos distinguir entre la correlación de Pearson y de Spearman, al igual que mostrar las correlaciones en formato de matriz.

EDA

Por último, hemos creado un **análisis exploratorio** (EDA) desde la recogida de los datos, el cálculo de las tablas de frecuencia y todas las métricas asociadas, principales gráficos y conclusiones que se pueden extraer.

¡Enhorabuena! Fastbook superado

edix

Creamos Digital Workers