**Análisis de Datos: Dispersión de variables numéricas**

Para saber si los datos están muy dispersos o se concentran en torno a un valor (caso en el que una de las medidas centrales como la media o la mediana los representarán muy bien), hemos visto ya los percentiles y el rango para hacernos una idea, pero en general lo que se emplea en el caso de los valores numéricos son dos medidas, relacionadas entre sí, y luego métodos visuales basados en dos tipos de gráficas: los histogramas y las gráficas de función de densidad de probabilidad [Aunque a esta parte le dedicaremos la siguiente píldora]

Vamos a verlos y a aplicarlos a nuestros dos casos de uso de ejemplo.

**import** numpy **as** np

**import** matplotlib.pyplot **as** plt

**import** pandas **as** pd

**import** seaborn **as** sns

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

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

**Medidas de dispersión o variabilidad: Varianza**

La [varianza](https://es.wikipedia.org/wiki/Varianza) es la media aritmética del cuadrado de las desviaciones respecto a la media de un conjunto de datos (por ejemplo, los valores de una de nuestras variables o columnas y, en general, una distribución estadística). La varianza intenta describir la dispersión de los [*datos*](https://es.wikipedia.org/wiki/Dato)*. Básicamente representa lo que varían los datos*. **Como está elevada al cuadrado, la varianza no puede tener las mismas unidades que los datos**.

Una varianza elevada significa que los datos están más dispersos. Mientras que un valor bajo, indica que los datos están próximos a la media.

Texto

Descripción generada automáticamente con confianza baja

INCISO: Y por qué está elevado al cuadrado te preguntarás: Porque no quiero que las diferencias positivas y negativas se compensen. Piensa en esta serie de datos:

[-104,100,102,0,-100,120,-119]

La media es:

Imagen que contiene Diagrama

Descripción generada automáticamente

Si no eleváramos al cuadrado, el denominador sería:



Y al dividirlo se nos quedaría en un grado de dispersión de 0.28 que no es real (fíjate que el rango como tal es 239), para que no se compensen las diferencias se elevan al cuadrado.

Y la varianza según la definición de ese conjunto de datos es:



En vez de aplicarla a los datos veamos la versión comparable (es decir medida en las mismas unidades que los datos que estamos analizando) que es la desviación estándar.

**Medidas de dispersión o variabilidad: Desviación estándar**

La [desviación típica](https://es.wikipedia.org/wiki/Desviaci%C3%B3n_t%C3%ADpica) es la raíz cuadrada de la varianza.

Imagen que contiene Diagrama

Descripción generada automáticamente

Una ventaja que presenta la desviación estándar sobre la varianza es que se expresa en unidades de la variable en cuestión.

¿Y para qué nos sirven estas medidas?

* Nos dan una idea rápida de si los datos están dispersos (gte. compararemos la desviación con la media) y por tanto de si la media y mediana son buenos representantes de los valores o bien tenemos que trabajar la variable de otra forma, considerando rangos por ejemplo.
* Nos sirven para hacer cálculos posteriores y de otras medidas (que nos permitirán interpretar los datos de otras formas)

Podemos aplicar directamente el concepto de "Coeficiente de Variación" (CV) que es la división de la desviación estándar entre la media. Como pautas generales:

* Un CV menor al 15% suele considerarse como una baja variabilidad.
* Un CV entre 15% y 30% indica una variabilidad moderada.
* Un CV mayor al 30% a menudo se considera como una alta variabilidad.

Estos valores son orientativos y deben interpretarse en el contexto específico de tus datos y el área de estudio.

**Caso 1. Seguros: Dispersión**

En general fijate en la desviación y de nuevo la podemos obtener del método describe():

df\_seguros.describe()

Interfaz de usuario gráfica

Descripción generada automáticamente

Si obtenemos el coeficiente de variación:

**def** variabilidad(df):

df\_var**=**df.describe().loc[["std","mean"]].T

df\_var["CV"]**=**df\_var["std"]**/**df\_var["mean"]

**return** df\_var

variabilidad(df\_seguros)

Tabla

Descripción generada automáticamente

Conclusión (no para mostrar sino para seguir avanzando en el EDA):

* Salvo quizá monthly\_premium\_alto para el resto de las variables debo analizar con cuidado esa distribución de valores.
* En el caso de CLV que es una de nuestras directoras, y en el de income que es importante, tendré que dar más cariño

**Caso 2. Viajes: Dispersión**

Aplicando ya directamente la función:

variabilidad(df\_air\_jun)

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Nos ocurre algo parecido, en este caso probablemente sería más interesante hacerlo por compañía todo, pero eso lo veremos en la siguiente unidad. En cualquier caso, tanto consumo (CV > 90%) como ingresos, nuestras variables directoras (o por lo menos las que nosotros hemos decidido que lo sean por ahora) necesitan que miremos sus distribuciones.

En ambos casos tenemos que ver qué "pinta" tienen los datos, no nos vale con estas pistas de las métricas y ahí es donde entran nuestras dos siguientes herramientas: Los histogramas y las funciones de densidad.