TIC

Lección 2: Carga de datos y análisis descriptivo

Al realizar el análisis de un conjunto de datos con alguna herramienta de analítica (Python + pandas) es común necesitar una descripción de los datos, ya sea de un conjunto de datos en total, de las columnas, de las filas y otros que permitan conocer cómo están estructurados los datos y qué información se encuentra modelada por esos datos.

El análisis descriptivo de datos es el proceso de examinar y resumir un conjunto de datos para obtener una visión comprensible y significativa de los mismos. A través de las técnicas de análisis descriptivo, podemos obtener información relevante sobre la distribución, tendencia central, dispersión, relación entre variables y otros aspectos clave de los datos. El objetivo principal del análisis descriptivo de datos es proporcionar una descripción completa y precisa de los datos, de manera que podamos comprender su estructura y características principales. Esto nos ayuda a detectar patrones, identificar posibles valores atípicos y resumir la información. Es el primer paso cuando se inicia cualquier tarea de ciencia de los datos. Para realizar el análisis descriptivo de una base de datos, se suelen separar las variables que el conjunto de datos tiene. Posteriormente se encuentra la descripción estadística de cada una de las variables calculando mediciones como:

- Media de los datos
- Mediana Moda
- Rango Desviación
- estándar Varianza

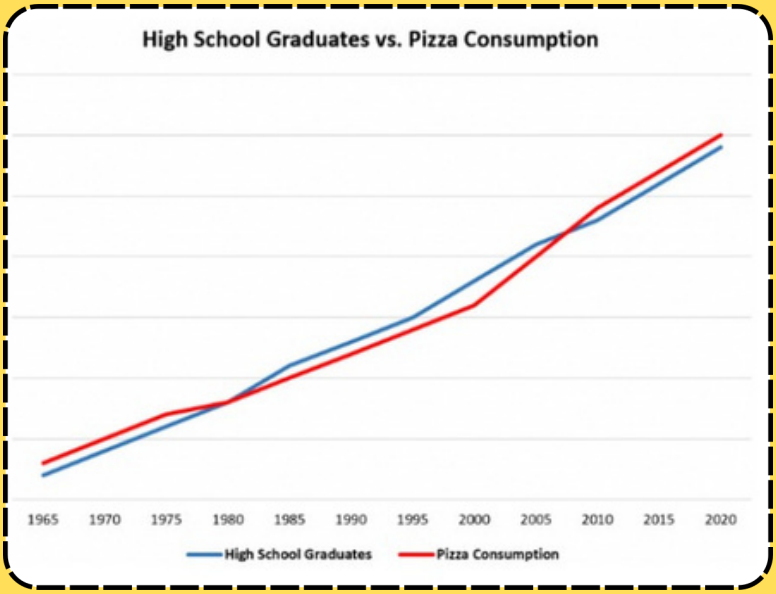
Con esto se sabe cómo se comportan las variables aleatorias que hacen parte de los datos. El siguiente paso es visualizar el comportamiento de las variables. A lo largo del curso se han realizado algunos de estos pasos de análisis descriptivo empleando Excel. Algunos gráficos para la exploración de datos son:

- Histogramas
- Diagramas de dispersión
- Gráficos de cajas y
- Bigotes Gráficos de barras.

Un paso adicional que se suele realizar es revisar cómo las variables aleatorias en el conjunto de datos están relacionadas entre si. Esto se lleva a cabo empleado un análisis de correlación. La correlación permite saber qué tanto afecta el cambio de una variable a otra. Si las alteraciones de una variable modifican otra, se dice que están correlacionadas. Para medir qué tanto se correlacionan, se emplea el coeficiente de correlación de Pearson.

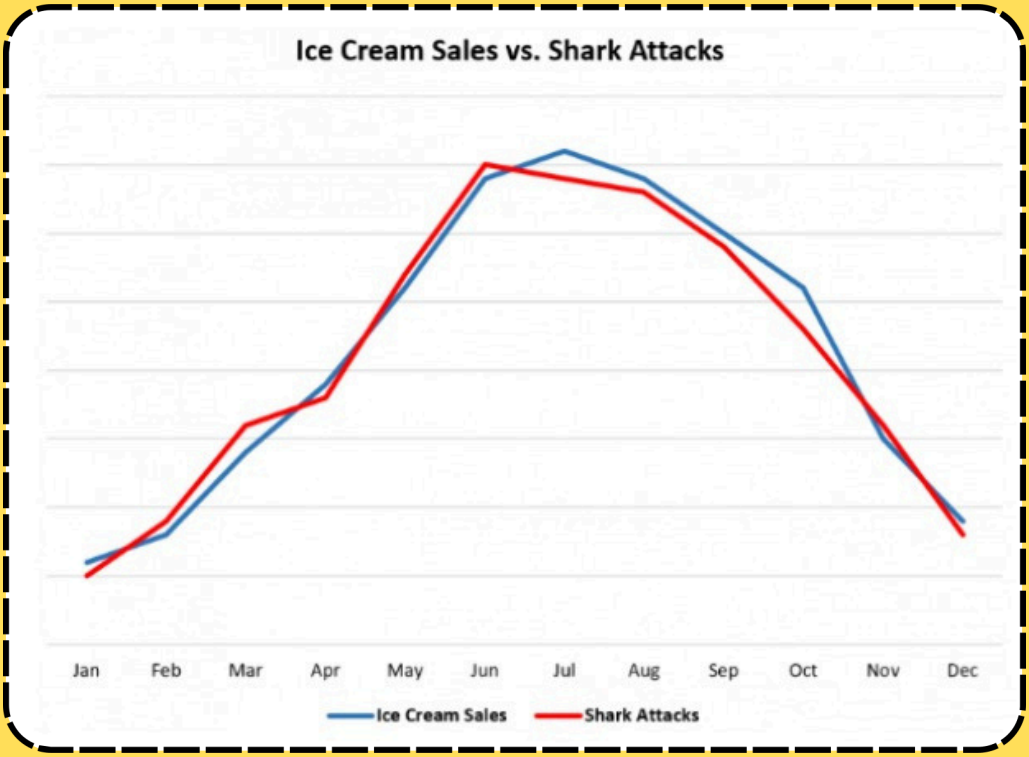
Nota importante

La correlación no implica causalidad. Esto es, el hecho de que dos variables sean similares o el comportamiento parezca similar (están correlacionadas) no quiere decir que una cause a la otra o viceversa. Una forma graciosa de entenderlo es mediante el análisis de algunas variables que muestran alta correlación, sin embargo, no tienen nada que ver una con otra. Por ejemplo: en estados unidos se analizaron las ventas de helado y se encontró que están correlacionadas con los ataques de tiburón:



Si bien están correlacionados, una cosa no causa la otra. Los ejemplos citados son exagerados, sin embargo, es común que los analistas de datos menos experimentados suelan pensar que existen datos que generan otros cuando el coeficiente de correlación calculado es alto. Una vez realizados los cálculos descritos lo más importante es poder entender qué datos hay en el conjunto de datos, qué datos son los más representativos, qué análisis se pueden aplicar y que modelos se pueden emplear sobre el conjunto de datos, entendiendo su distribución. Y cómo se relacionan las variables que contiene el conjunto de datos.

Como se aprecia, las variables están altamente correlacionadas, sin embargo, nada tiene que ver una cosa con la otra. Otro ejemplo es la comparación entre los graduados de la escuela secundaria y el consumo de pizza:



Si bien están correlacionados, una cosa no causa la otra. Los ejemplos citados son exagerados, sin embargo, es común que los analistas de datos menos experimentados suelan pensar que existen datos que generan otros cuando el coeficiente de correlación calculado es alto. Una vez realizados los cálculos descritos lo más importante es poder entender qué datos hay en el conjunto de datos, qué datos son los más representativos, qué análisis se pueden aplicar y que modelos se pueden emplear sobre el conjunto de datos, entendiendo su distribución. Y cómo se relacionan las variables que contiene el conjunto de datos.