Universidad Tecnológica de Puebla

División: **Tecnologías de la Información**

Ingeniería en Desarrollo y Gestión de Software

Asignatura:

Extracción de conocimiento en bases de datos

Método Describe()

Docente:

María del Rosario Sánchez Bañuelos

Alcantara Flores Christian

Grupo: **9°B**

Cuatrimestre: **Mayo – Agosto 2022**

Describe()

La función describe de Pandas permite mostrar datos estadísticos de una forma entendible “coloquialmente”, puesto que con la información que se pasa en el archivo .csv, consigue obtener datos numéricos que considera importantes para cuantificar y detallar, permitiendo obtener los valores de media, mediana, valores máximos, mínimos y conteos para una columna en particular de los datos.

Esta función solo retorna valores cuando los datos son puramente numéricos.

```
In [44]: datos.describe()
```

Out[44]:

	SNo	Confirmed	Deaths	Recovered
count	3395.000000	3395.000000	3395.000000	3395.000000
mean	1698.000000	611.823859	17.756112	167.704271
std	980.196409	5121.319656	187.195366	1650.055341
min	1.000000	0.000000	0.000000	0.000000
25%	849.500000	2.000000	0.000000	0.000000
50%	1698.000000	10.000000	0.000000	1.000000
75%	2546.500000	120.500000	1.000000	18.000000
max	3395.000000	67332.000000	2871.000000	38557.000000

- Los **datos obtenidos** en el dataframe muestran una cantidad de registros de **3395,000000**
- **En datos percentiles** se muestra una partición de la contabilización total que se considera en la primera fila, esto para los registros de **25%, 50%, 75%.** Un ejemplo de esto es que en la muestra del 25% indica **Sno:849,50000**, que multiplicado por 4 **25*5=10,** obtiene el conteo general de la primera fila de: **3395.00000**.
- Mean, representa el valor de la mediana, un valor intermedio para los datos que están ordenados según la librería o la función de Describe.

- La desviación estándar esta dada por la fila std: Esta desviación mide la dispersión de una distribución de datos. Entre más dispersa está una distribución de datos, más grande es su desviación.

Información de los datos – Método **.info()**

```
In [43]: datos.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3395 entries, 0 to 3394
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   SNo              3395 non-null   int64
 1   ObservationDate  3395 non-null   object
 2   Province/State   2264 non-null   object
 3   Country/Region   3395 non-null   object
 4   Last Update      3395 non-null   object
 5   Confirmed        3395 non-null   float64
 6   Deaths           3395 non-null   float64
 7   Recovered        3395 non-null   float64
dtypes: float64(3), int64(1), object(4)
memory usage: 212.3+ KB
```

Este método recoge parte de las columnas importantes de nuestro archivo de datos .csv y realiza una lectura sobre los registros que se tienen de cada uno de estos, obteniendo el nombre de la columna, y contenido generado, además de especificar el tipo de dato que maneja, esto por ser lenguaje no tipado representa de manera genérica los objetos. También este método obtiene datos de la cantidad de almacenamiento que ocupa la información.