Guía básica para la limpieza y el reconocimiento de datos

Pasar de texto a columnas un archivo en formato CSV.

1. Seleccionar solapa datos
2. Seleccionar columna.
3. Clic en texto en columnas.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

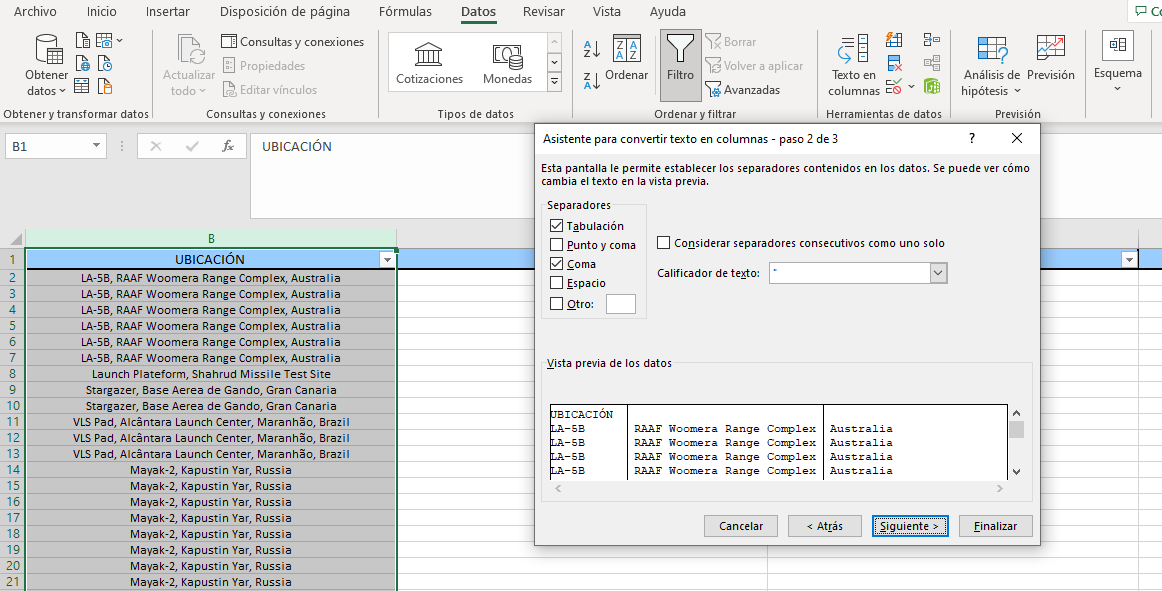
Interfaz de usuario gráfica, Texto, Aplicación

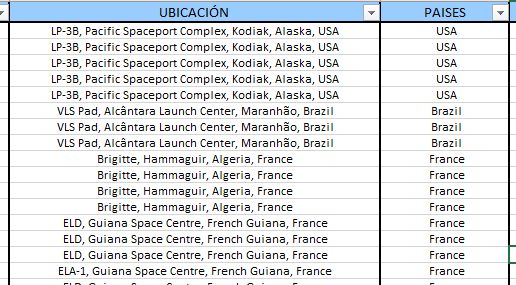
Descripción generada automáticamente

Tabla

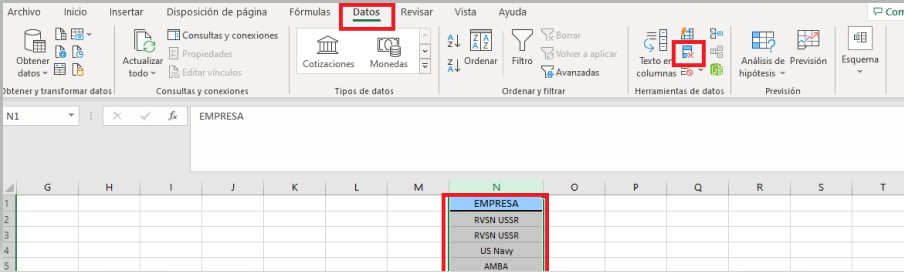
Descripción generada automáticamente con confianza media

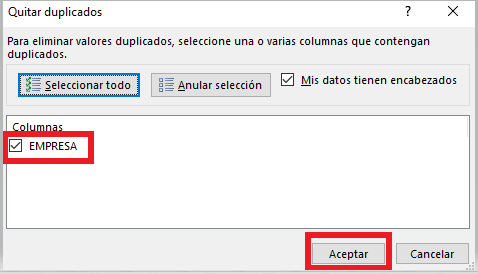
Otra opción para usar texto en columna

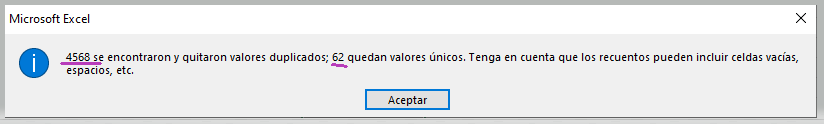




Eliminar duplicados



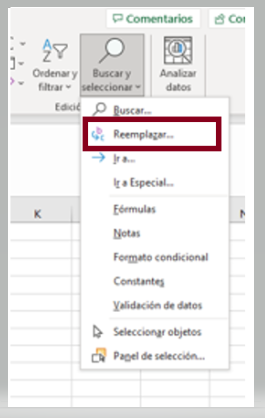


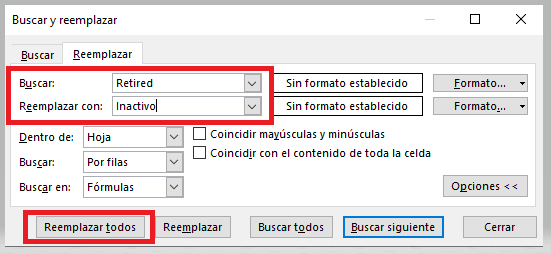


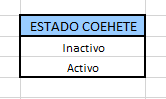
Reemplazar valores

Tabla

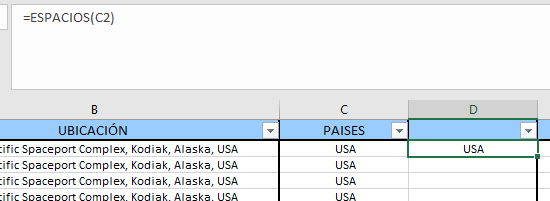
Descripción generada automáticamente







Función espacio (TRIM)



Funciones para extraer texto de una columna

En Excel, hay varias funciones que puedes utilizar para extraer texto de una columna o celda. Aquí te muestro algunas de las funciones más comunes:

* **Izquierda:** Extrae un número específico de caracteres desde el principio (izquierda) de una cadena de texto.

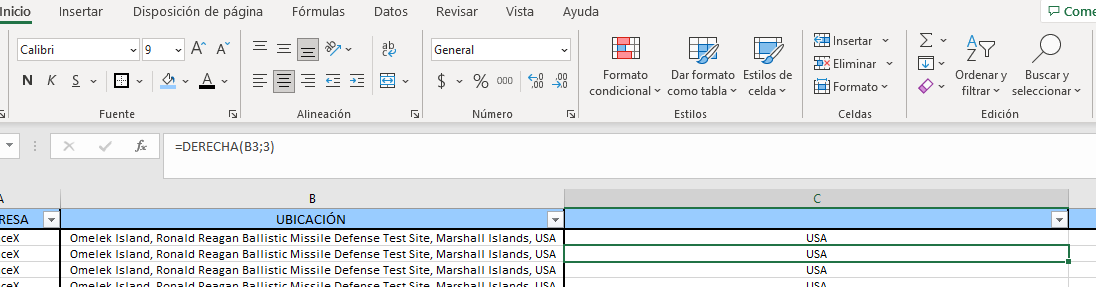
Ejemplo: =IZQUIERDA(A1, 5) extraerá los primeros 5 caracteres de la celda A1.

Tabla

Descripción generada automáticamente

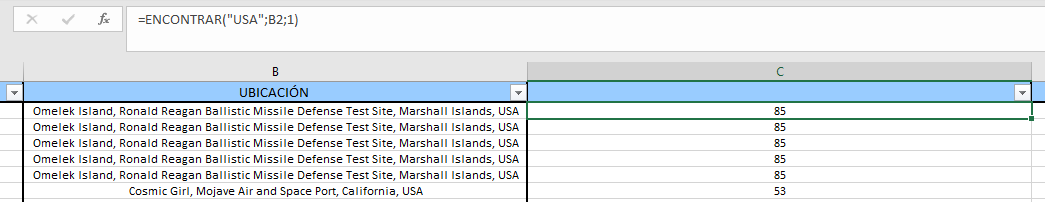
* **Derecha:** Extrae un número específico de caracteres desde el final (derecha) de una cadena de texto.

Ejemplo: =DERECHA(A1, 3) extraerá los últimos 3 caracteres de la celda A1.



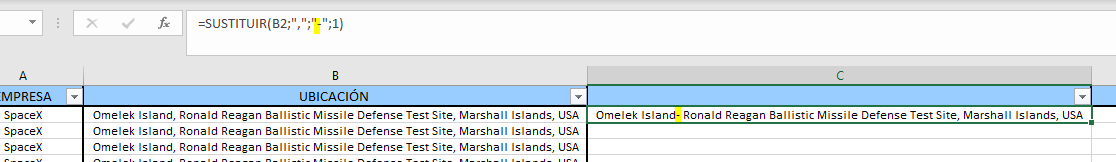
* **Encontrar:** Busca un texto específico dentro de una cadena y devuelve la posición donde se encuentra la primera ocurrencia.

Ejemplo: =ENCONTRAR(",", A1) encontrará la posición de la primera coma en la celda A1.



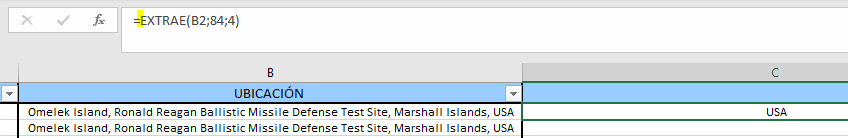
* **Sustituir:** Reemplaza un texto específico dentro de una cadena por otro texto.

Ejemplo: =SUSTITUIR(A1, " ", "-") reemplazará todos los espacios en blanco por guiones en la celda A1.

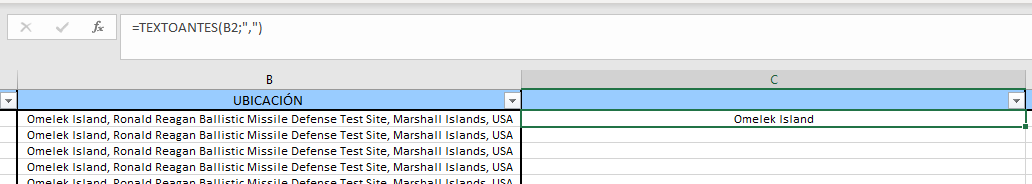


* **Extrae:** Extrae un fragmento de texto de una cadena usando un número de inicio y un número de caracteres.

Ejemplo: =EXTRAE(A1, 5, 8) extraerá 8 caracteres desde la posición 5 en la celda A1.



* **Texto antes:**

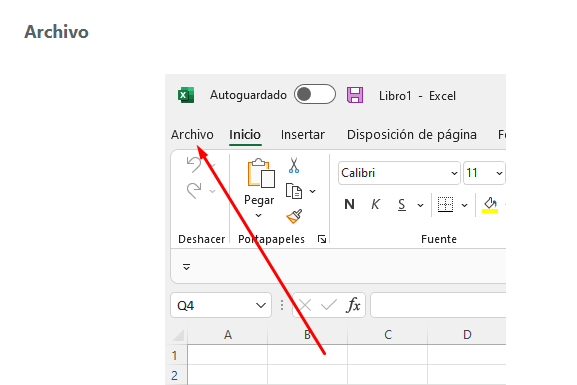


* **Texto después:**

Interfaz de usuario gráfica, Aplicación, Tabla

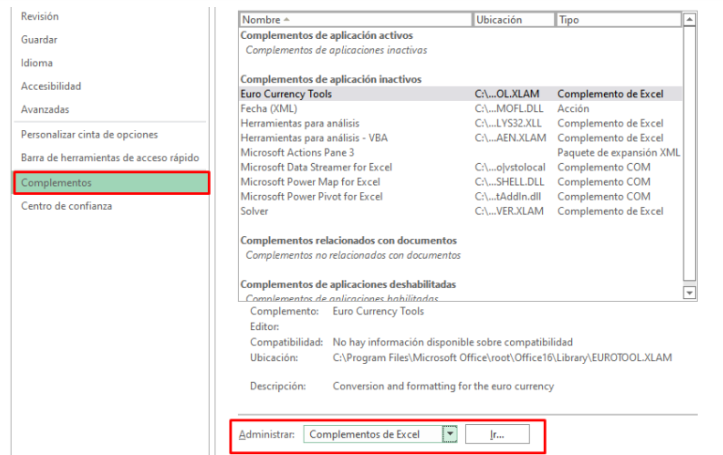
Descripción generada automáticamente

Habilitar la Herramienta de Análisis de Datos en Excel



Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

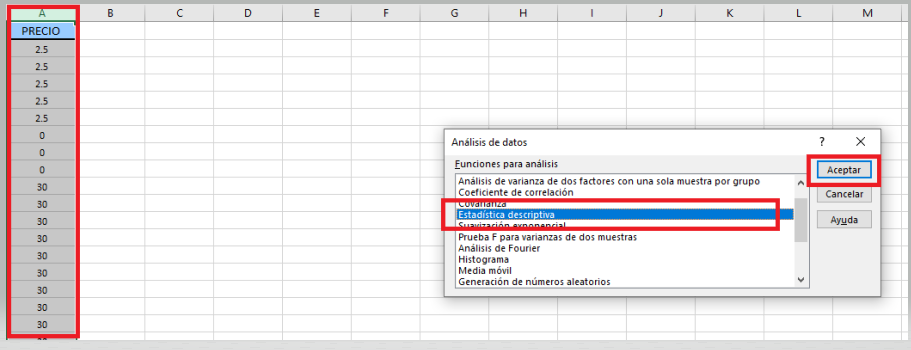


Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación, Tabla, Excel

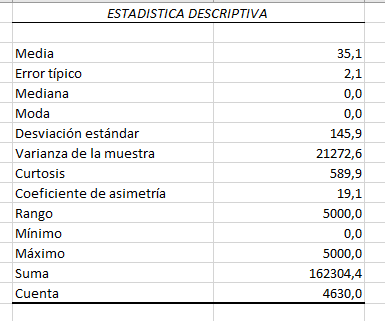
Descripción generada automáticamente



Interfaz de usuario gráfica, Aplicación, Tabla, Excel

Descripción generada automáticamente

Estadística descriptiva



**Media**

la media es simplemente el promedio de un conjunto de números. Para calcularla, sumamos todos los valores y luego dividimos esa suma entre la cantidad de valores que tenemos. Es útil para obtener una idea general de qué tan grande o pequeño es un conjunto de datos.

* Interpretación: Si tenemos las edades de un grupo de personas y la media es 30 años, eso significa que, en promedio, las personas del grupo tienen 30 años.

**Error típico**

El error típico (también conocido como desviación estándar de la media) nos dice cuánto varían los datos respecto a la media. Es una medida de dispersión.

Interpretación: Un error típico pequeño indica que los datos están más cercanos a la media y son más consistentes. Un error típico grande indica que los datos están más dispersos y pueden variar mucho.

**Mediana**

La mediana es el valor que está justo en el medio de un conjunto de datos ordenados. Si tenemos un número impar de valores, la mediana es el valor central. Si tenemos un número par de valores, es el promedio de los dos valores centrales.

* Interpretación: Si tenemos una lista de salarios y encontramos que la mediana es $25,000, eso significa que la mitad de los salarios están por encima de $25,000 y la otra mitad está por debajo.

**Moda**

La moda es el valor que aparece con mayor frecuencia en un conjunto de datos. Puede haber una, ninguna o varias modas en un conjunto de datos.

* Interpretación: Si tenemos una lista de calificaciones de un examen y la moda es 90, eso significa que el puntaje más común obtenido en el examen fue 90.

**Desviación estándar**

La desviación estándar es otra medida de dispersión que nos indica cuánto se alejan los datos de la media. Cuanto mayor sea la desviación estándar, mayor será la dispersión de los datos.

* Interpretación: Si tenemos un conjunto de datos de alturas y encontramos que la desviación estándar es 5 cm, eso significa que las alturas de las personas tienden a variar alrededor de 5 cm respecto a la media.

**Varianza de la muestra**

La varianza de la muestra es una medida de dispersión similar a la desviación estándar. Es el promedio de los cuadrados de las diferencias entre cada dato y la media.

* Interpretación: Una varianza grande indica que los datos están más dispersos respecto a la media, mientras que una varianza pequeña indica que están más cercanos a la media.

**Curtosis**

La curtosis mide qué tan puntiaguda o achatada es la distribución de datos en relación con una distribución normal. Puede ser leptocúrtica (puntiaguda), mesocúrtica (similar a una distribución normal) o platicúrtica (achatada).

* Interpretación: La curtosis nos ayuda a entender cómo se distribuyen los datos en torno a la media. Una curtosis positiva indica una distribución más puntiaguda, mientras que una curtosis negativa indica una distribución más achatada.

**Coeficiente de asimetría**

El coeficiente de asimetría mide la asimetría de la distribución de datos, es decir, si se inclina hacia la izquierda o hacia la derecha. Un coeficiente positivo indica asimetría hacia la derecha, mientras que uno negativo indica asimetría hacia la izquierda.

* Interpretación: Si tenemos un conjunto de datos de ingresos y el coeficiente de asimetría es positivo, eso significa que hay más datos con ingresos bajos y unos pocos con ingresos muy altos.

**Rango**

El rango es simplemente la diferencia entre el valor máximo y el valor mínimo de un conjunto de datos. Es una medida de la amplitud del conjunto.

* Interpretación: Si tenemos una lista de temperaturas máximas durante una semana y el rango es 20°C, eso significa que la diferencia entre la temperatura máxima más alta y la más baja es de 20°C.

**Mínimo**

El mínimo es el valor más pequeño en un conjunto de datos.

**Máximo**

El máximo es el valor más grande en un conjunto de datos.

**Suma**

La suma simplemente es el resultado de agregar todos los valores de un conjunto.

**Cuenta**

La cuenta nos dice cuántos valores hay en un conjunto de datos.

**Conclusión del ejemplo:**

Para interpretar los resultados de la estadística descriptiva, es importante analizar cada valor y lo que representa en el contexto del conjunto de datos. A continuación, te proporciono algunas conclusiones basadas en los resultados que has proporcionado:

**Media**: La media es 35.1. Esto indica que, en promedio, los datos tienden a estar alrededor de este valor.

**Error típico**: El error típico es 2.1. Indica que los datos pueden variar alrededor de la media en aproximadamente 2.1 unidades.

**Mediana**: La mediana es 0.0. Esto significa que la mitad de los datos están por encima de cero y la otra mitad están por debajo. La mediana es mucho más baja que la media, lo que indica que puede haber valores extremadamente altos que están afectando la media.

**Moda**: La moda es 0.0. Esto indica que el valor más común en el conjunto de datos es cero.

**Desviación estándar**: La desviación estándar es 145.9. Indica que los datos tienen una alta dispersión alrededor de la media. Hay valores que varían considerablemente de la media.

**Varianza de la muestra**: La varianza de la muestra es 21272.6. Es el promedio de las diferencias al cuadrado entre cada valor y la media. Una varianza alta también indica que los datos tienen mucha variabilidad.

**Curtosis**: La curtosis es 589.9. Un valor muy alto de curtosis indica una distribución con colas pesadas y valores extremadamente altos o bajos en comparación con una distribución normal.

**Coeficiente de asimetría**: El coeficiente de asimetría es 19.1. Un valor tan alto de coeficiente de asimetría indica una asimetría extrema hacia la derecha, lo que significa que hay valores atípicos muy altos que están sesgando la distribución hacia ese lado.

**Rango**: El rango es 5000.0. Indica la diferencia entre el valor máximo y el valor mínimo en el conjunto de datos.

**Mínimo**: El valor mínimo es 0.0. Es el valor más pequeño en el conjunto de datos.

**Máximo**: El valor máximo es 5000.0. Es el valor más grande en el conjunto de datos.

**Suma**: La suma total de los datos es 162304.4.

**Cuenta**: Hay un total de 4630 datos en el conjunto.

En general, estos resultados indican que el conjunto de datos tiene una alta variabilidad, con valores que varían significativamente alrededor de la media y una distribución asimétrica hacia la derecha debido a la presencia de valores atípicos muy altos. Es importante examinar los datos con más detalle para entender la naturaleza de estos valores extremos y su posible impacto en el análisis estadístico.

Tabla Pivot

**Gráficos**

1. Empresas por geografía.

| **Etiquetas de fila** | **Cuenta de EMPRESA** |
| --- | --- |
| Australia | 6 |
| Barents Sea | 3 |
| Brazil | 3 |
| China | 354 |
| France | 318 |
| Gran Canaria | 2 |
| India | 82 |
| Iran | 16 |
| Israel | 11 |
| Japan | 123 |
| Kazakhstan | 719 |
| Kenya | 9 |
| New Zealand | 28 |
| North Korea | 5 |
| Pacific Missile Range Facility | 1 |
| Pacific Ocean | 36 |
| Russia | 1411 |
| South Korea | 5 |
| USA | 1467 |
| China | 11 |
| Japan | 7 |
| Russia | 5 |
| Shahrud Missile Test Site | 1 |
| USA | 4 |
| Yellow Sea | 3 |
| **Total general** | **4630** |

1. Misiones por Estado

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente



1. Cohetes por Estado

Interfaz de usuario gráfica, Aplicación, Tabla, Word

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente