

Data Preparation con ChatCSV



por Lucas Castronuovo



¿Qué es Data Preparation?

1 Preparación de los datos

Proceso en el cual los datos sin procesar se transforman para su posterior análisis

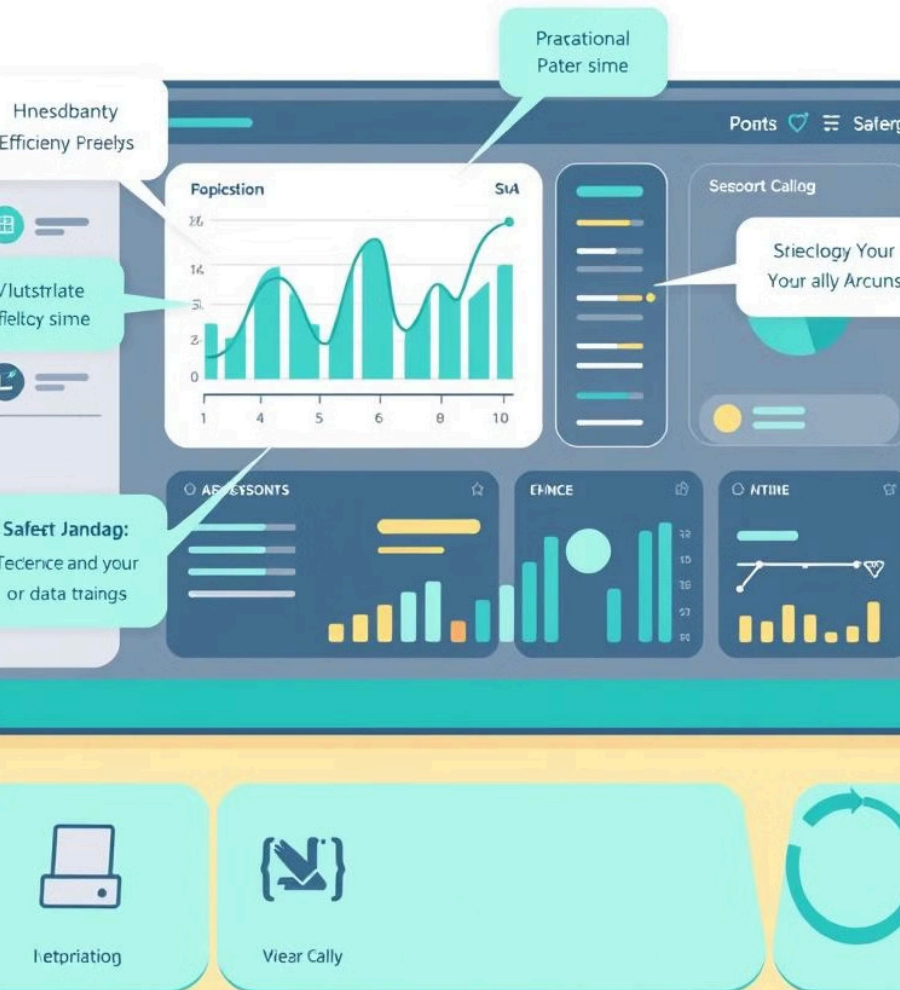
2 Su importancia

Es un componente especial especial para los proyectos de Machine Learning

3 Tareas clave

- Recopilación
- Limpieza
- Etiquetado
- Exploración
- Visualización

Características de sus herramientas



Detección rápida de errores

Se corrigen los problemas evitando complicaciones posteriores

Producción de datos de alta calidad

Datos limpiados y reformateados, asegurando que se generen resultados óptimos y precisos

Mejora en la toma de decisiones

Como resultado, se llegan a mejores conclusiones y decisiones

Alivio en la carga de trabajo

Permite que los esfuerzos humanos se inviertan en las tareas estratégicas

ChatCSV

Desarrollada por AirOps

Empresa que se especializa en crear herramientas de inteligencia artificial y automatización diseñadas para facilitar el análisis y procesamiento de datos

Innovación en el análisis de datos

Mantiene una conversación con el usuario implementando el archivo CSV como elemento interactivo

Características principales

- Interacción conversacional única
- Sugerencias de consultas
- Visualización de datos
- Historial de chats
- Soporte y preguntas frecuentes

**Advantages it ussrg sing
data prepatation tool.**



Ventajas de usar ChatCSV

1

Colaboración en tiempo real

2

Paneles personalizables

3

Seguridad en el manejo de datos

4

Soporte Multilingüe

5

**Compatibilidad con Dispositivos
Móviles**

Cons of using a
balta preparen tool.



Contras de usar ChatCSV

1

Curva de Aprendizaje

2

Preocupaciones de Privacidad

3

Limitación de Tipos de Archivo

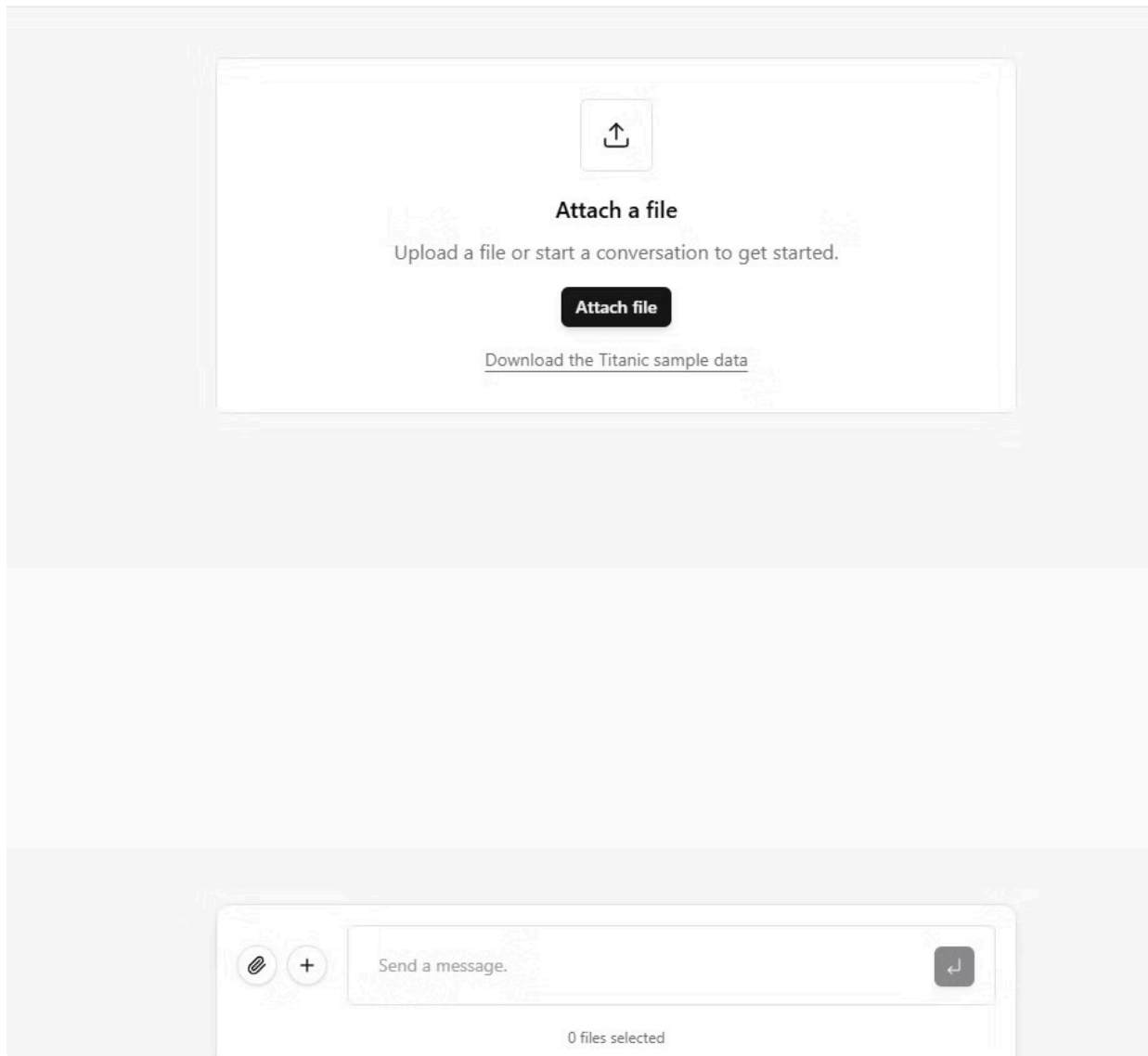
4

Dependencia de Conexión a Internet

5

Visualizaciones Genéricas

Implementación de la herramienta





Give me a quick rundown of this data

Loading large file...

The dataframe df1 has 149,116 entries and 11 columns. Here is a brief description of each column:

1. `transaction_id`: This is an integer column and it does not contain any null values. It likely represents a unique identifier for each transaction.
2. `transaction_date`: This is an object column and it does not contain any null values. It likely represents the date of each transaction.
3. `transaction_time`: This is an object column and it does not contain any null values. It likely represents the time of each transaction.
4. `transaction_qty`: This is an integer column and it does not contain any null values. It likely represents the quantity of items in each transaction.
5. `store_id`: This is an integer column and it does not contain any null values. It likely represents a unique identifier for each store.
6. `store_location`: This is an object column and it does not contain any null values. It likely represents the location of each store.
7. `product_id`: This is an integer column and it does not contain any null values. It likely represents a unique identifier for each product.

Send a message.

1 files selected

De qué trata el CSV?

Loading large file...

El CSV parece ser un registro de transacciones de una tienda o cadena de tiendas. Cada fila representa una transacción individual, con detalles sobre la fecha y hora de la transacción, la cantidad de productos comprados, la tienda en la que se realizó la compra, y detalles sobre el producto comprado.

Aquí hay un desglose de lo que cada columna parece representar:

1. ``transaction_id``: Un identificador único para cada transacción.
2. ``transaction_date``: La fecha en que se realizó la transacción.
3. ``transaction_time``: La hora en que se realizó la transacción.
4. ``transaction_qty``: La cantidad de productos comprados en la transacción.
5. ``store_id``: Un identificador único para cada tienda.
6. ``store_location``: La ubicación de la tienda donde se realizó la transacción.
7. ``product_id``: Un identificador único para cada producto.

Cuenta con valores nulos?

Loading large file...

No, el dataframe no tiene valores nulos. Cada columna en el dataframe tiene cero valores nulos.

How did I do?

Agrega al CSV la columna "periodo" donde tienes que categorizar cada registro si la transacción fue realizada a la "Mañana" (si la hora de transacción fue entre las 6:00 y las 11:59), a la "Tarde" (si la hora de transacción fue entre las 12:00 y las 17:59), o a la "Noche" (si la hora de transacción fue entre las 18:00 y las 5:59) (Fíjate en el horario de `transaction_time`). Luego tienes que agregar otra columna llamada temporada donde tendrás que categorizar cada transacción según la estación del año en que fue realizada (ten en cuenta el mes y si el día es 21 por el cambio de estación) (Fíjate en el día y mes de `transaction_date`). Elimina las columnas `transaction_time` y `transaction_date`

Aquí está el enlace al CSV actualizado: [updated_dataframe](#).

transaction_id	transaction_qty	store_id	stc
1	2	5	
2	2	5	
3	2	5	
4	1	5	
5	2	5	
6	1	5	
7	1	5	
8	2	5	
9	1	5	
10	2	5	
11	1	5	
12	2	5	
13	1	5	
14	1	5	
15	2	5	

Aquí está el enlace al CSV actualizado: [updated_dataframe](#).

product_detail	periodo	temporada
Ethiopia Rg	Mañana	Invierno
Spicy Eye Opener Chai ...	Mañana	Invierno
Dark chocolate Lg	Mañana	Invierno
Our Old Time Diner Ble...	Mañana	Invierno
Spicy Eye Opener Chai ...	Mañana	Invierno
Oatmeal Scone	Mañana	Invierno
Our Old Time Diner Ble...	Mañana	Invierno
Columbian Medium Roa...	Mañana	Invierno
Latte Rg	Mañana	Invierno
Dark chocolate Rg	Mañana	Invierno
Spicy Eye Opener Chai ...	Mañana	Invierno
Ethiopia Lg	Mañana	Invierno
Earl Grey Lg	Mañana	Invierno
Spicy Eye Opener Chai ...	Mañana	Invierno
Ouro Brasileiro shot	Mañana	Invierno





Elimina las columnas store_id product_id

Las columnas que no sean de tipo int o float conviértelas en valores numéricos (incluyendo fechas y horarios) Con LabelEncoder. Dime que valor le asignaste a cada categoría de las columnas "periodo" y "temporada"



Loading large file...

The LabelEncoder assigned the following values to the categories in the "periodo" and "temporada" columns:

For "periodo":

- 'Invierno': 0
- 'Primavera': 1
- 'Verano': 2

For "temporada":

- 'Invierno': 0
- 'Primavera': 1
- 'Verano': 2

Re-run response

Share



Send a message.



1 files selected



En periodo las categorías son "Mañana" "Tarde" y "Noche". Dime sus valores



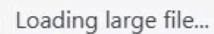
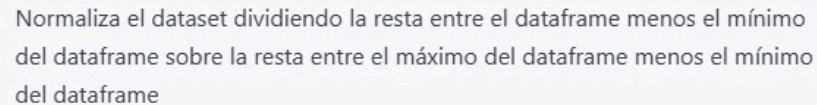
Loading large file...

The LabelEncoder assigned the following values to the categories in the "periodo" column:

- 'Mañana': 0
- 'Noche': 1
- 'Tarde': 2

How did I do?





I have normalized the dataframe by subtracting the minimum value of each column from each value in that column, and then dividing by the range of that column (the maximum value minus the minimum value). This scales all values in the dataframe to be between 0 and 1.

Now, I will save this normalized dataframe to a CSV file. I have saved the normalized dataframe to a CSV file. You can access it using the following link:

normalized dataset

action qty	store location	unit price	product c
85714285714285	1	0.0497737556561086	0.25
85714285714285	1	0.05203619909502262	1
85714285714285	1	0.083710407239819	0.5
	1	0.027149321266968323	0.25
85714285714285	1	0.05203619909502262	1
	1	0.0497737556561086	0
	1	266968323	0.25



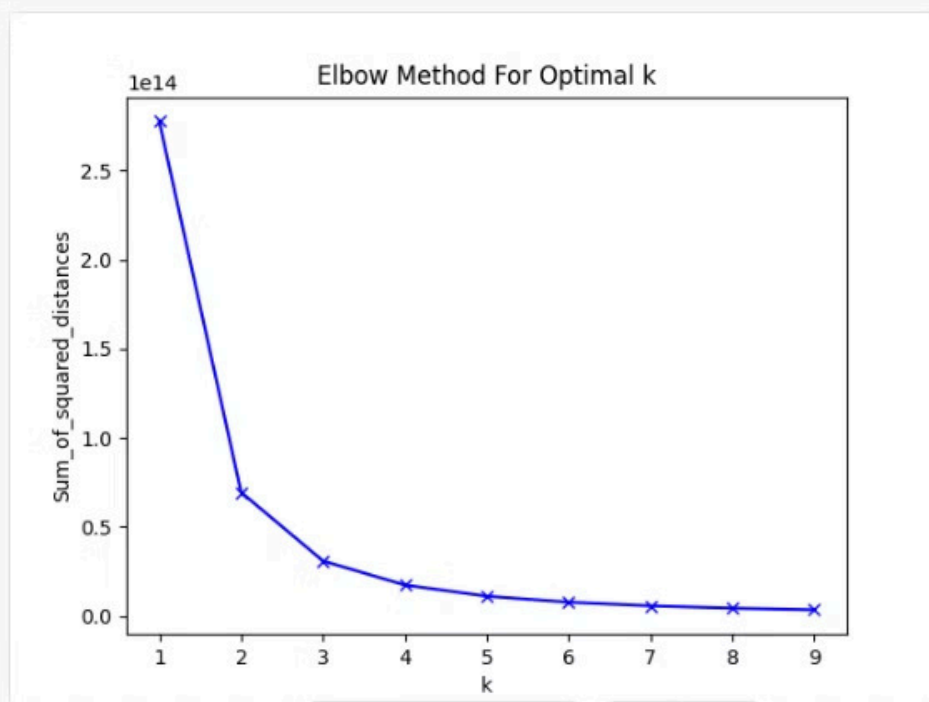


Grafica el "Codo de Jambú" para determinar la cantidad de clusters a utilizar.
Qué cantidad me recomiendas utilizar?



Loading large file...

Here is the Elbow Method plot:



Re-run response

Share

The Elbow Method is a heuristic used in determining the number of clusters in a data set. The idea of the elbow method is to run k-means clustering on the dataset for a range of values of k (say, k from 1 to 10), and for each value of k calculate the sum of squared errors (SSE).

Then, plot a line chart of the SSE for each value of k. If the line chart looks like an arm, then the "elbow" on the arm is the value of k that is the best. The idea is that we want a small SSE, but that the SSE tends to decrease toward 0 as we increase k (the SSE is 0 when k is equal to the number of data points in the dataset, because then each data point is its own cluster, and there is no error between it and the center of its cluster).

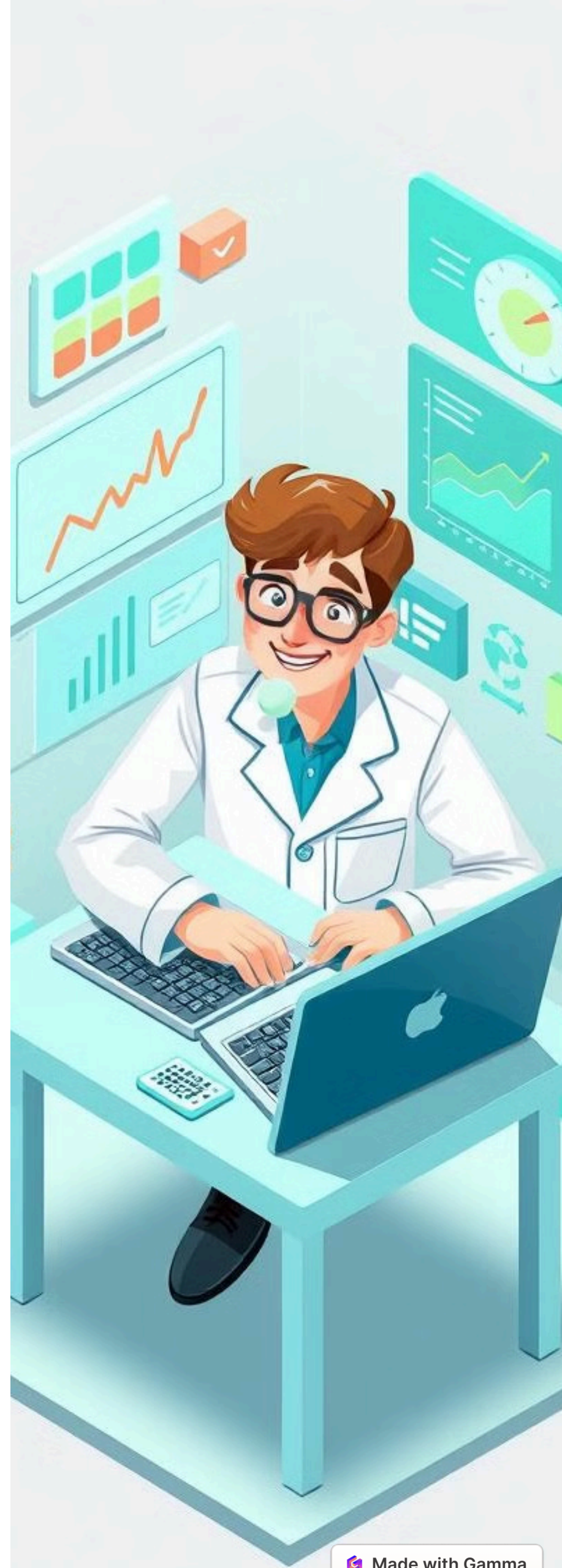
So our goal is to choose a small value of k that still has a low SSE, and the elbow usually represents where we start to have diminishing returns by increasing k.

From the plot, it seems like the elbow is around $k=3$. So, I would recommend using 3 clusters.

How did I do? 😊 😞

Re-run response

Share





Conclusiones

- 1 **Herramienta transformadora**
- 2 **Aún por mejorar**
- 3 **Futuro prometedor para estas herramientas**