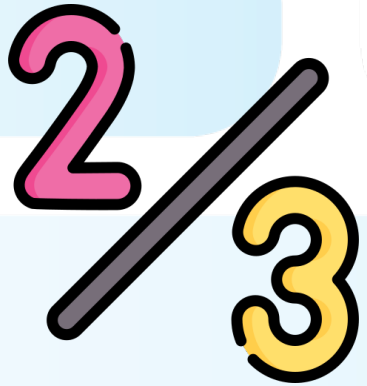
Consideraciones al crear subconjuntos

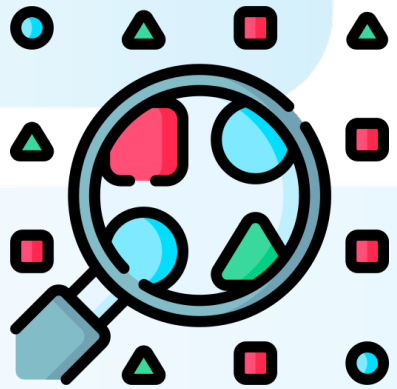
Consideraciones al crear subconjuntos



- Es importante dividir el dataset en al menos dos conjuntos
- Podemos usar la función popular `train_test_split`

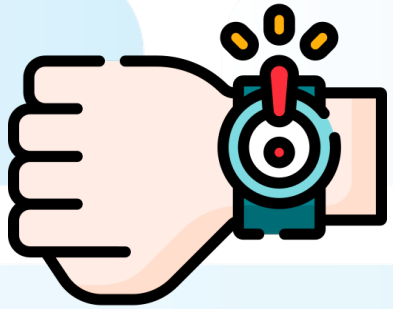
El problema con `train_test_split`

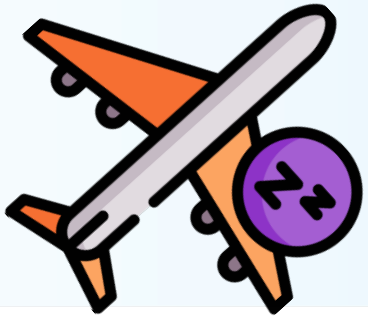
- La selección es aleatoria
- La aleatoriedad es no repetible
- Si tu intención es comparar dos modelos, divisiones similares deben ser usadas



Fair splitting

- Hay fenómenos que están relacionados entre ello
  - Órdenes de entrega en restaurantes
  - Vuelos en un aeropuerto
- Verifica la relación que tienen tus observaciones entre si



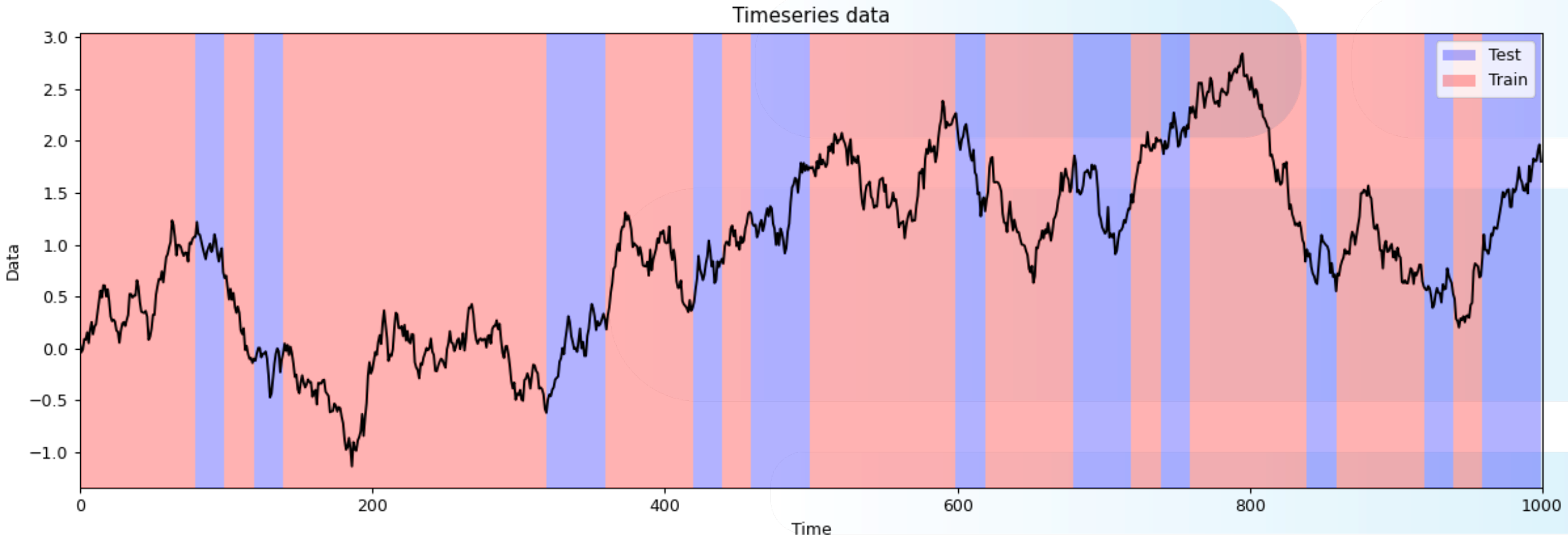


Series de tiempo

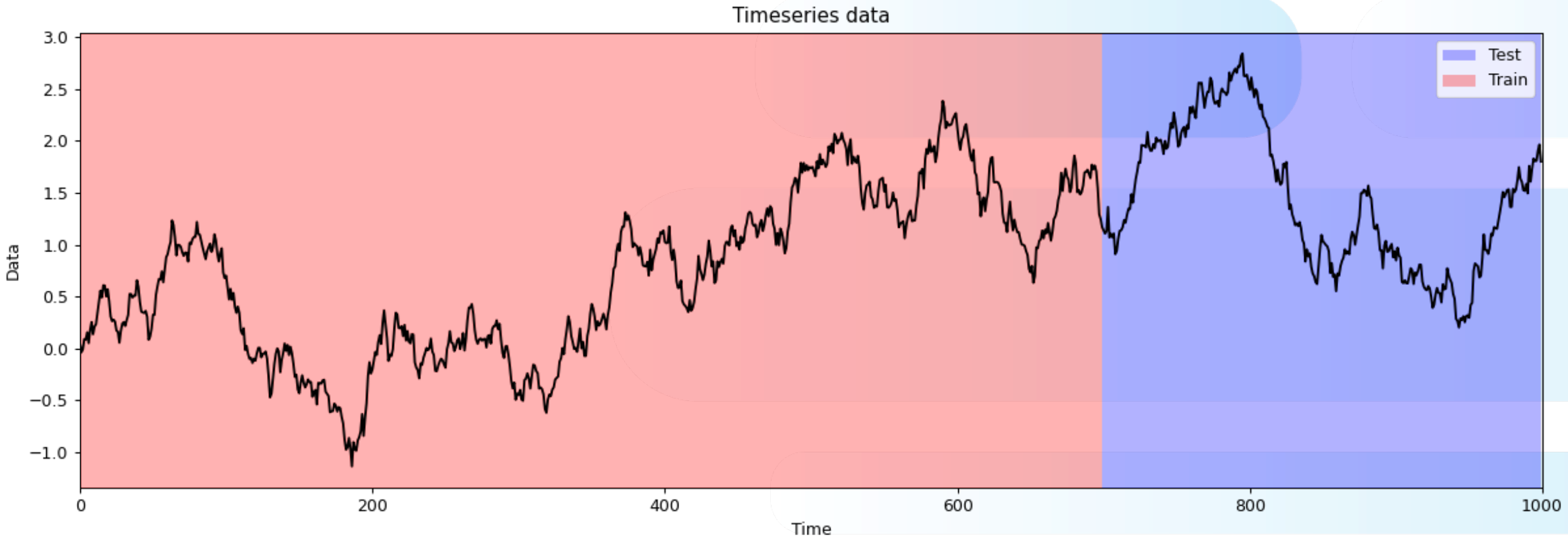
- Las observaciones pasadas tienen influencia directa a lo esperado en el futuro

- La división aleatoria no tiene en cuenta estas relaciones

Series de tiempo dividida aleatoriamente



Series de tiempo dividida justamente



Conclusión

- Asegúrate de que el spliting sea determinístico
- Prevén el data leakage en series de tiempo