# PD1 Ejer1:

## Data:

1. La porción más importante y pesada del trabajo es la limpieza y mezclado de datos
2. Blending es la unión y recolección de datos necesarios, cleansing es la limpieza de datos inútiles o erróneos/vacíos
3. Para manejar datos faltantes se puede: Reemplazar por otro valor (como el promedio, siempre y cuando no afecte gravemente las métricas que necesitemos), eliminar el atributo si tiene demasiados vacíos y no aporta mucho o solo eliminar los ejemplos que los tienen vacíos (si no hay muchos)
4. Se pueden desactivar módulos puntuales

## Outliers

1. Hay varios métodos de eliminar datos outliers
2. En general es necesario normalizarlos antes de poder detectarlos, en RM use el normalizador que usa estandarización (z-transform)
3. Se puede usar la distancia euclideana para distinguirlos, que nos deja definir que tanta distancia buscar y la cant de outliers que queremos sacar
4. El módulo detect outliers agrega un atributo booleano que dice si tal ejemplo es un outlier o no, el filtrado se debe hacer por fuera
5. Se pueden colocar breakpoints antes y después de cada modulo para revisar y debuggear los datos conforme se van transformando
6. Se pueden invertir los filtros fácilmente con un campo en los parámetros