**INTEGRANTES:**

- ✓ María Augusta Jimbo Granda
- ✓ María Magdalena Romero Guzmán

# PRÁCTICA N° 2

## 1. Descripción del dataset. Por qué es importante y qué pregunta/problema pretende responder?

El dataset que vamos a utilizar en la presente práctica es el de clasificación de barras de chocolate a nivel mundial, es decir; el conjunto de datos llamado "flavors_of_cacao.csv"

**Contexto:**

El chocolate es uno de los dulces más populares en el mundo. Cada año, los residentes de los Estados Unidos colectivamente comen más de 2.8 billones de libras. Sin embargo, ¡no todas las barras de chocolate son iguales! Este conjunto de datos contiene calificaciones de expertos de más de 1.700 barras de chocolate individuales, junto con información sobre su origen regional, el porcentaje de cacao, la variedad de alubia de chocolate utilizada y dónde se cultivaron los frijoles.

Sistema de calificación de sabores de cacao:
5 = Elite (que trasciende los límites normales)
4 = Premium (Desarrollo de sabor superior, carácter y estilo)
3 = Satisfactorio (3.0) a loable (3.75) (bien hecho con cualidades especiales)
2 = decepcionante (pasable pero contiene al menos un defecto significativo)
1 = Desagradable (principalmente desagradable)

La base de datos se enfoca principalmente en el chocolate negro simple, con el objetivo de apreciar los sabores del cacao cuando se lo convierte en chocolate. Las calificaciones no reflejan los beneficios de salud, las misiones sociales o el estado orgánico.

La diversidad, el equilibrio, la intensidad y la pureza de los sabores son considerados. La genética, el terruño, las técnicas de poscosecha, el procesamiento y el almacenamiento pueden discutirse cuando se considera el componente de sabor.

La textura tiene un gran impacto en la experiencia general y también es posible que los problemas relacionados con la textura afecten el sabor. Es una buena forma de evaluar la visión del fabricante, la atención al detalle y el nivel de competencia.

El dataset contiene 1795 registros sin incluir el encabezado y 9 columnas o atributos que son:

| # | Nombre | Descripción | Tipo |
|---|---|---|---|
| 1 | Empresa | Nombre de la empresa que fabrica la barra | String |
| 2 | Geo-región | La geo región de origen específica para la barra | String |
| 3 | REF | Un valor vinculado a cuando se ingresó la revisión en la base de datos. Más alto = más reciente | Numeric |
| 4 | Fecha de revisión | Fecha de publicación de la revisión | Numeric |
| 5 | Porcentaje de cacao | Porcentaje de cacao (oscuridad) de la barra de chocolate que se revisa | Numeric |
| 6 | Localización | País base del fabricante | String |
| 7 | Calificación | Calificación de expertos | Numeric |
| 8 | Tipo de frijol | La variedad de frijol utilizada, en el caso de que se proporcione | String |
| 9 | Origen de haba | La amplia geo región de origen para el haba | String |

Las preguntas principales que se pueden revisar, analizar y evaluar son, por ejemplo:

- ¿Dónde se cultivan los mejores granos de cacao?
- ¿Qué países producen las barras mejor calificadas?
- ¿Cuál es la relación entre el porcentaje de sólidos de cacao y la calificación?

## 2. Integración y selección de los datos de interés a analizar.

Los datos que se analizarán son:

| # | Nombre | Descripción | Tipo |
|---|---|---|---|
| 1 | Empresa | Nombre de la empresa que fabrica la barra | String |
| 2 | Geo-región | La geo región de origen específica para la barra | String |
| 3 | Porcentaje de cacao | Porcentaje de cacao (oscuridad) de la barra de chocolate que se revisa | Numeric |
| 4 | Localización | País base del fabricante | String |
| 5 | Calificación | Calificación de expertos | Numeric |

Lo anterior, en razón de que parecen ser los más importantes al momento de obtener las conclusiones.

## 3. Limpieza de los datos.

### 3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Los datos si contienen elementos vacíos.

Gestionaríamos cada uno de estos mediante el uso de funciones propias del lenguaje R Studio, al momento de construir el algoritmo correspondiente.

### 3.2. Identificación y tratamiento de valores extremos.

Los valores extremos son aquellos que parecen no ser congruentes si los comparamos con el resto de los datos.

Los valores extremos pueden causar serios problemas para los análisis estadísticos. Primero, generalmente sirven para aumentar la varianza del error y reducir el poder de las pruebas estadísticas. En segundo lugar, si no se distribuye aleatoriamente, pueden alterar sustancialmente las probabilidades de cometer errores tipo I y tipo II. En tercer lugar, pueden sesgar o influir seriamente en las estimaciones que pueden ser de interés sustancial porque pueden no ser generadas por la población de interés.

Por lo antes indicado, de la revisión del conjunto de datos en forma preliminar, se visualiza que estos son consistentes y están dentro de rangos normales por cada uno de los atributos, es por ello que el manejo de estos valores extremos consistirá en simplemente dejarlos como actualmente están levantados, es decir; no sufren ningún tipo de alteración.