ENTREGA DEL PROYECTO FINAL

Definición fuente de datos

Fuente de datos – Kaggle – Coffee Quality Database from CQI https://
www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi

1. Descripción de la temática de los datos

Para el presente trabajo se obtuvo la información de la evaluación y cata de diversas cosechas de café realizados por distintos examinadores de la empresa certificadora internacional Coffee Quality Institute a cosechas de café de 36 distintos países. Se obtuvieron dos bases de datos, una correspondiente a la variedad robusta y otra a la variedad arábica de cafés. Cada base presenta información relacionada a las medidas de calidad de la taza de café como Aroma, Sabor, Acidez, Equilibrio, etc. También presenta información sobre el procesamiento del grano, la localidad de cosecha, entre otros.

Se presentan dos bases de datos que contienen reseñas de evaluadores especializados para ambos tipos de café: arábica y robusta, respectivamente. La información anterior se proporciona en cada conjunto de datos. Las primeras cosechas evaluadas se remontan al 9 de abril del 2010 y el horizonte de las evaluaciones se extiende hasta el 11 de enero del 2018.

2. Hipótesis

Comprender si características de la localidad de cosecha, procesamiento y variedad de los granos de café afecta la puntuación final de la taza de café. Para esto se evaluará las puntuaciones otorgadas a los cafés durante toda la historia de evaluaciones de CQI y comprobaremos si ciertas combinaciones de las características de los granos conllevan consistentemente a una mayor puntuación. Después comprobaremos que las dimensiones evaluadas en cada taza de café como aroma, sabor, acidez, entre otras, son consistentemente afectadas por las características de la cosecha del grano de café.

Al saber qué características de la cosecha del café y procesamiento conlleva a un mayor puntaje en la certificación del café se podría generar recomendaciones a las empresas cafetaleras respecto a la zona donde desarrollen sus cosechas para optimizar las características del café que quieren obtener y la puntuación mayor que deberá conllevar a mayores ventas y menores costos de experimentación si la empresa desea crear un café con un perfil de taza específico.

3. Datasets

A continuación, se adjunta los csv de las bases de datos organizadas y estructuradas utilizadas en el presente trabajo:

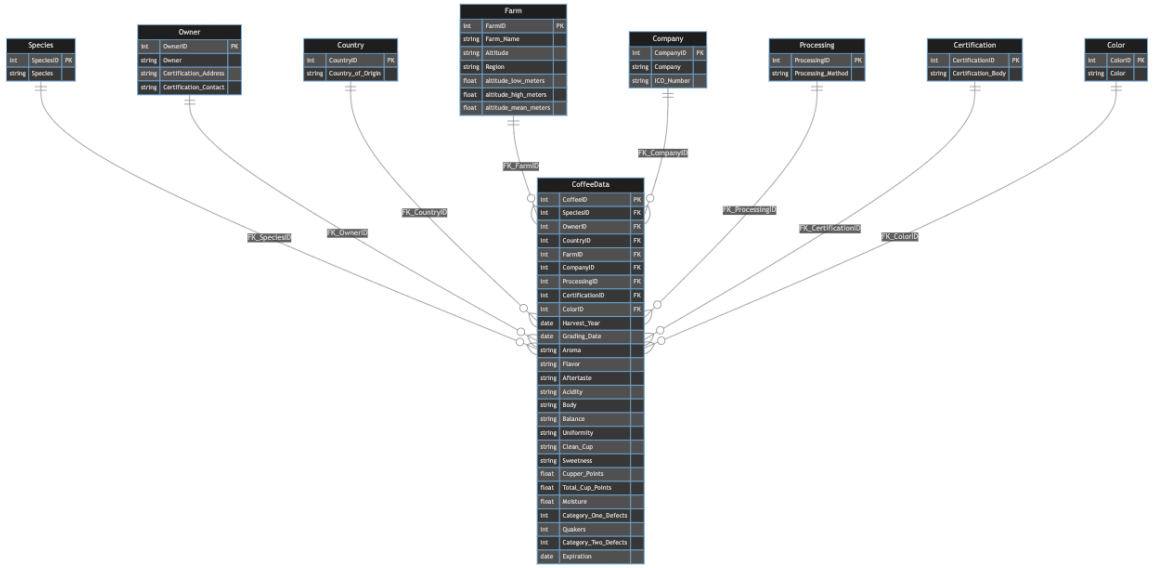
Base de Datos - Tablas de evaluaciones de café.zip

CODERHOUSE

ENTREGA DEL PROYECTO FINAL

Definición fuente de datos

4. Diagrama de Entidad-Relación



5. Hipótesis

5. Listado de Tablas

En este apartado se hará mención de cada una de las tablas junto a una breve descripción de las mismas y la definición de las claves primaria y foráneas:

Species: Tipo de especie
- Primaria: SpeciesID (Tipo de especie)

Owner: Nombre del propietario
- Primaria: OwnerID (Nombre del propietario)

Country: País de origen
- Primaria: CountryID (País de origen)

Farm: Nombre de la finca
- Primaria: FarmID (Nombre de la finca)

Company: Nombre de la empresa
- Primaria: CompanyID (Nombre de la empresa)

Processing: Método de procesamiento
- Primaria: ProcessingID (Método de procesamiento)

Certification: Cuerpo de certificación
- Primaria: CertificationID (Cuerpo de certificación)

Color: Color del grano
- Primaria: ColorID (Color del grano)

Coffee Data: Datos detallados del café
- Foránea: SpeciesID (Datos detallados del café)
- Foránea: OwnerID (ID de especie)
- Foránea: CountryID (ID de propietario)
- Foránea: FarmID (ID de país)
- Foránea: CompanyID (ID de finca)
- Foránea: ProcessingID (ID de empresa)
- Foránea: CertificationID (ID de procesamiento)
- Foránea: ColorID (ID de certificación)
- Primaria: CoffeeID (ID de color)

CODERHOUSE

ENTREGA DEL PROYECTO FINAL

6. Transformaciones Realizadas y Medidas Calculadas

Transformaciones de Datos:

1. Conversión del Peso de la Bolsa a Kilogramos (`Bag.Weight`):
   - Se convierte el peso de la bolsa a kilogramos, considerando las unidades 'kg', 'lbs' y 'lb'.

Formula DAX:
```
Weight in KG =
IF(
    SEARCH("kg", data[Bag.Weight], 1, BLANK()) <> BLANK(),
    VALUE(REPLACE(data[Bag.Weight], SEARCH("kg", data[Bag.Weight]), 2, "")),
    IF(
        SEARCH("lbs", data[Bag.Weight], 1, BLANK()) <> BLANK(),
        VALUE(REPLACE(data[Bag.Weight], SEARCH("lbs", data[Bag.Weight]), 3, "")) * 0.453592,
        VALUE(REPLACE(data[Bag.Weight], SEARCH("lb", data[Bag.Weight]), 2, "")) * 0.453592
    )
)
```

Funciones DAX equivalentes:
- `SEARCH()`: Para buscar subcadenas dentro de una cadena.
- `REPLACE()`: Para reemplazar subcadenas dentro de una cadena.
- `VALUE()`: Para convertir una cadena a un número.

2. Conversión de las Columnas de Fecha a Formato Consistente (YYYY-MM-DD):
   - Las columnas `Grading.Date` y `Expiration` se convierten al formato de fecha YYYY-MM-DD.

Formula DAX:
```
Grading Date = FORMAT(data[Grading.Date], "YYYY-MM-DD")
Expiration Date = FORMAT(data[Expiration], "YYYY-MM-DD")
```

Funciones DAX equivalentes:
- `FORMAT()`: Para formatear una fecha en el formato deseado.

3. Cálculo de la Diferencia de Años (`Year_Difference`):
   - Se calcula la diferencia entre el año de vencimiento y el año de cosecha. Al ver que la correlación es casi 1 se utiliza un factor de aproximadamente 1 año de diferencia para imputar los valores faltantes de

Fórmula DAX:
```
Year Difference = YEAR(data[Expiration]) - YEAR(data[Harvest.Year])
```

Funciones DAX equivalentes:
- `YEAR()`: Para extraer el año de una fecha.
- Operaciones aritméticas básicas para calcular la diferencia.

4. Manejo de Datos Faltantes:
   - Se imputan valores faltantes en varias columnas, como `Owner`, `Country.of.Origin`, `Region`, `Variety`, `Processing.Method`, entre otras. En el caso particular de valores en blanco en `Country.of.Origin` se imputa el valor de 'Colombia' luego de revisar otros proyectos que trabajaron con esta data y descubrieron que este era el valor omitido para este campo.

Fórmula DAX:
```
Owner = IF(ISBLANK(data[Owner]), data[Farm.Name], data[Owner])
Country of Origin = IF(ISBLANK(data[Country.of.Origin]), "Colombia", data[Country.of.Origin])
Region = IF(ISBLANK(data[Region]), data[Country.of.Origin], data[Region])
```

Funciones DAX equivalentes:
- `IF()`: Para condiciones.
- `ISBLANK()`: Para verificar si un valor está vacío.
- `COALESCE()`: Para devolver el primer valor no vacío en una lista.

5. Corrección de Datos:
   - Se corrige el valor de `Country.of.Origin` si es 'United States' usando el valor de `Region`

Funciones DAX equivalentes:
- `IF()`: Para condiciones.
- `ISBLANK()`: Para verificar si un valor está vacío.

Fórmula DAX:
```
Country of Origin =
IF(
    data[Country.of.Origin] = "United States",
    data[Region],
    data[Country.of.Origin]
)
```

ENTREGA DEL PROYECTO FINAL

Definición fuente de datos

7. Médidas Calculadas

1. Diferencia Media (`mean_diff`):
   - Se calcula la diferencia media entre el año de vencimiento y el año de cosecha.

   Fórmula DAX:
   Mean Difference = AVERAGE(data[Year_Difference])

2. Diferencia Mediana (`median_diff`):
   - Se calcula la diferencia mediana entre el año de vencimiento y el año de cosecha.

   Fórmula DAX:
   Median Difference = MEDIAN(data[Year_Difference])

3. Cuartiles (`quantiles_diff`):
   - Se calculan los cuartiles de la diferencia entre el año de vencimiento y el año de cosecha.

   Fórmula DAX:
   Q1 Difference = PERCENTILE.INC(data[Year_Difference], 0.25)
   Q2 Difference = PERCENTILE.INC(data[Year_Difference], 0.5)
   Q3 Difference = PERCENTILE.INC(data[Year_Difference], 0.75)

4. Desviación Estándar (`std_diff`):
   - Se calcula la desviación estándar de la diferencia entre el año de vencimiento y el año de cosecha.

   Fórmula DAX:
   Std Dev Difference = STDEV.P(data[Year_Difference])

5. Correlación (`correlation`):
   - Se calcula la correlación entre `Harvest.Year` y `Expiration_Year`.

   Fórmula DAX:
   Correlation = CORRELATION(data[Harvest.Year], data[Expiration_Year])

ENTREGA DEL PROYECTO FINAL

Definición fuente de datos

7. Tablas Calendario

A continuación presento un breve detalle sobre las tablas de calendario creadas y las funciones DAX de PowerBI que fuero utilizadas en función de las columnas de fecha disponibles en el dataset que son las correspondientes a la fecha de calificación y fecha de expiración.
Las columnas generadas en esta tabla son:

- Año (`Year`)
- Trimestre (`Quarter`)
- Mes (`Month`)
- Día (`Day`)
- Día de la semana (`Weekday`)

Utilizando las fórmulas DAX correspondientes de:

- `YEAR()`: Para extraer el año de la fecha.
- `QUARTER()`: Para determinar el trimestre de la fecha.
- `MONTH()`: Para extraer el mes de la fecha.
- `DAY()`: Para extraer el día de la fecha.
- `WEEKDAY()`: Para determinar el día de la semana.

Las tablas de calendario son esenciales para realizar análisis temporales en PowerBI.