
Calidad de Cafés del Mundo
Pablo Illescas

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

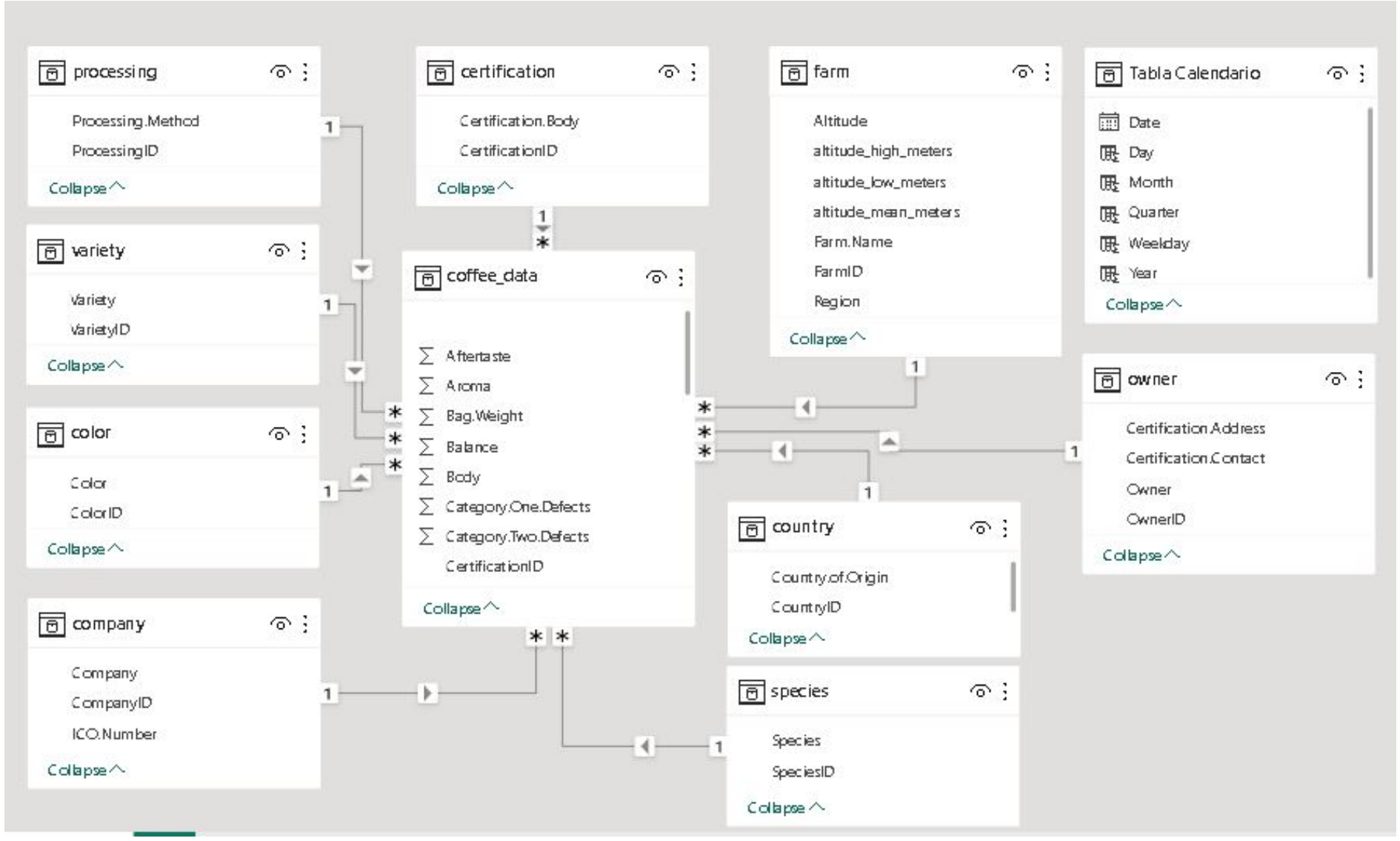
A continuación, se adjunta los csv de las bases de datos organizadas y estructuradas utilizadas en el presente trabajo:

Base de Datos - Tablas de evaluaciones de café.zip

CODERHOUSE

ENTREGA DEL PROYECTO FINAL

4. Diagrama de Entidad-Relación



ENTREGA DEL PROYECTO FINAL

Definición fuente de datos

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)

Variety: Tipo de variedad o sub especie
- Primaria: VarietyID (Tipo de variedad o sub especie)

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)

ENTREGA DEL PROYECTO FINAL

6. 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
- NK: Species (Tipo de especie)
- Primaria: SpeciesID (ID único para especies)

Owner: Nombre del propietario
- NK: Owner (Nombre del propietario)
- NK: Certification.Address (Dirección de certificación)
- NK: Certification.Contact (Contacto de certificación)
- Primaria: OwnerID (ID único para propietario)

Country: País de origen
- NK: Country.of.Origin (País de origen)
- Primaria: CountryID (ID único para país)

Farm: Nombre de la finca
- NK: Farm.Name (Nombre de la finca)
- NK: Altitude (Altitud)
- NK: Region (Región)
- NK: altitude_low_meters (Altitud baja en metros)
- NK: altitude_high_meters (Altitud alta en metros)
- NK: altitude_mean_meters (Altitud media en metros)
- Primaria: FarmID (ID único para finca)

Company: Nombre de la empresa
- NK: Company (Nombre de la empresa)
- NK: ICO.Number (Número ICO)
- Primaria: CompanyID (ID único para empresa)

Processing: Método de procesamiento
- NK: Processing.Method (Método de procesamiento)
- Primaria: ProcessingID (ID único para método de procesamiento)

Certification: Cuerpo de certificación
- NK: Certification.Body (Cuerpo de certificación)
- Primaria: CertificationID (ID único para cuerpo de certificación)

Color: Color del grano
- NK: Color (Color del grano)
- Primaria: ColorID (ID único para color)

Coffee Data: ID de especie
- Foránea: SpeciesID (ID de especie)
- Foránea: OwnerID (ID de propietario)
- Foránea: CountryID (ID de país)
- Foránea: FarmID (ID de finca)
- Foránea: CompanyID (ID de empresa)
- Foránea: ProcessingID (ID de procesamiento)
- Foránea: CertificationID (ID de certificación)
- Foránea: ColorID (ID de color)
- NK: Aroma (Año de cosecha)
- NK: Flavor (Fecha de clasificación)
- NK: Aftertaste (Aroma)
- NK: Acidity (Sabor)
- NK: Body (Regusto)
- NK: Balance (Acidez)
- NK: Uniformity (Cuerpo)
- NK: Clean.Cup (Equilibrio)
- NK: Sweetness (Uniformidad)
- NK: Cupper.Points (Taza limpia)
- NK: Total.Cup.Points (Dulzura)
- NK: Moisture (Puntos de catador)
- NK: Category.One.Defects (Puntos totales de taza)
- NK: Quakers (Humedad)
- NK: Category.Two.Defects (Defectos de categoría uno)
- NK: Expiration (Quakers)
- Primaria: CoffeeID (Defectos de categoría dos)

ENTREGA DEL PROYECTO FINAL

6. 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 (donde NK denota que un campo no es llave):

Species: Tipo de especie
- NK (string): Species (Tipo de especie)
- Primaria: SpeciesID (ID único para especies)

Owner: Nombre del propietario
- NK (string): Owner (Nombre del propietario)
- NK (string): Certification.Address (Dirección de certificación)
- NK (string): Certification.Contact (Contacto de certificación)
- Primaria: OwnerID (ID único para propietario)

Country: País de origen
- NK (string): Country.of.Origin (País de origen)
- Primaria: CountryID (ID único para país)

Farm: Nombre de la finca
- NK (string): Farm.Name (Nombre de la finca)
- NK (float): Altitude (Altitud)
- NK (string): Region (Región)
- NK (float): altitude_low_meters (Altitud baja en metros)
- NK (float): altitude_high_meters (Altitud alta en metros)
- NK (float): altitude_mean_meters (Altitud media en metros)
- Primaria: FarmID (ID único para finca)

Company: Nombre de la empresa
- NK (string): Company (Nombre de la empresa)
- NK (string): ICO.Number (Número ICO)
- Primaria: CompanyID (ID único para empresa)

Processing: Método de procesamiento
- NK (string): Processing.Method (Método de procesamiento)
- Primaria: ProcessingID (ID único para método de procesamiento)

Certification: Cuerpo de certificación
- NK (string): Certification.Body (Cuerpo de certificación)
- Primaria: CertificationID (ID único para cuerpo de certificación)

Color: Color del grano
- NK (string): Color (Color del grano)
- Primaria: ColorID (ID único para color)

Coffee Data: ID de especie
- Foránea: SpeciesID (ID de especie)
- Foránea: OwnerID (ID de propietario)
- Foránea: CountryID (ID de país)
- Foránea: FarmID (ID de finca)
- Foránea: CompanyID (ID de empresa)
- Foránea: ProcessingID (ID de procesamiento)
- Foránea: CertificationID (ID de certificación)
- Foránea: ColorID (ID de color)
- NK (int): Harvest.Year (Año de cosecha)
- NK (date): Grading.Date (Fecha de clasificación)
- NK (int): Aroma (Aroma)
- NK (int): Flavor (Sabor)
- NK (int): Aftertaste (Regusto)
- NK (int): Acidity (Acidez)
- NK (int): Body (Cuerpo)
- NK (int): Balance (Equilibrio)
- NK (int): Uniformity (Uniformidad)
- NK (int): Clean.Cup (Taza limpia)
- NK (int): Sweetness (Dulzura)
- NK (int): Cupper.Points (Puntos de catador)
- NK (int): Total.Cup.Points (Puntos totales de taza)
- NK (int): Moisture (Humedad)
- NK (int): Category.One.Defects (Defectos de categoría uno)
- NK (int): Quakers (Quakers)
- NK (int): Category.Two.Defects (Defectos de categoría dos)
- NK (date): Expiration (Fecha de vencimiento)
- Primaria: CoffeeID (ID único para datos de café)

ENTREGA DEL PROYECTO FINAL

7. 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

8. 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

9. 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`)

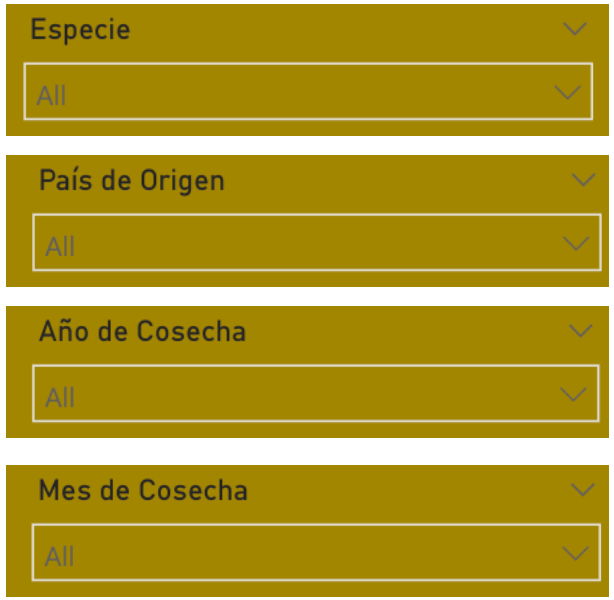
Utilizando las fórmulas DAX correspondientes de:

- `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.

10. Segmentaciones elegidas

En el reporte se seleccionaron las segmentaciones por filtros desplegables en las dimensiones de Especie, País de Origen, Año de Cosecha y Mes de Cosecha.



CODERHOUSE

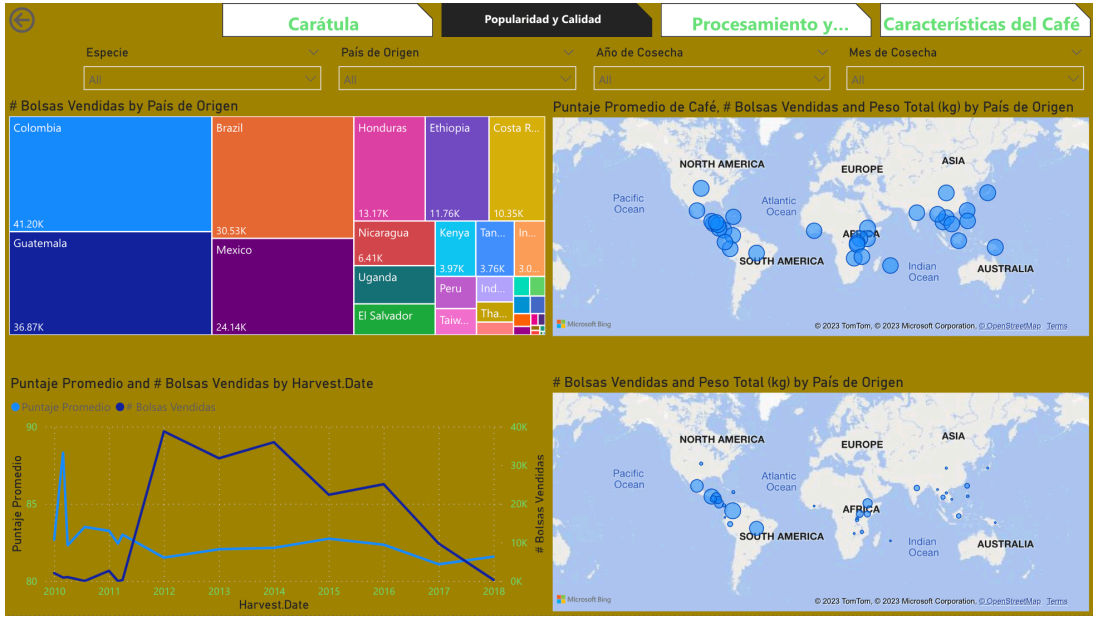
ENTREGA DEL PROYECTO FINAL

Visualización de los datos

11. Listado de páginas del reporte

El reporte en PowerBI está conformado por 4 páginas. La primera corresponde a una carátula que introduce la temática del reporte y contiene pestañas de navegación hacia el resto del reporte permitiendo una navegación lineal o a elección del usuario según el foco de interés.

Las otras páginas consisten en conjuntos de visualizaciones organizadas temáticamente cada una con un objetivo particular de presentación y que en conjunto soportan el objetivo del reporte que es proporcionar un vistazo de lo más general y descriptivo a lo más especifico y explicativo de la calidad del café y los factores que impactan en la calidad de la taza de café.



CODERHOUSE

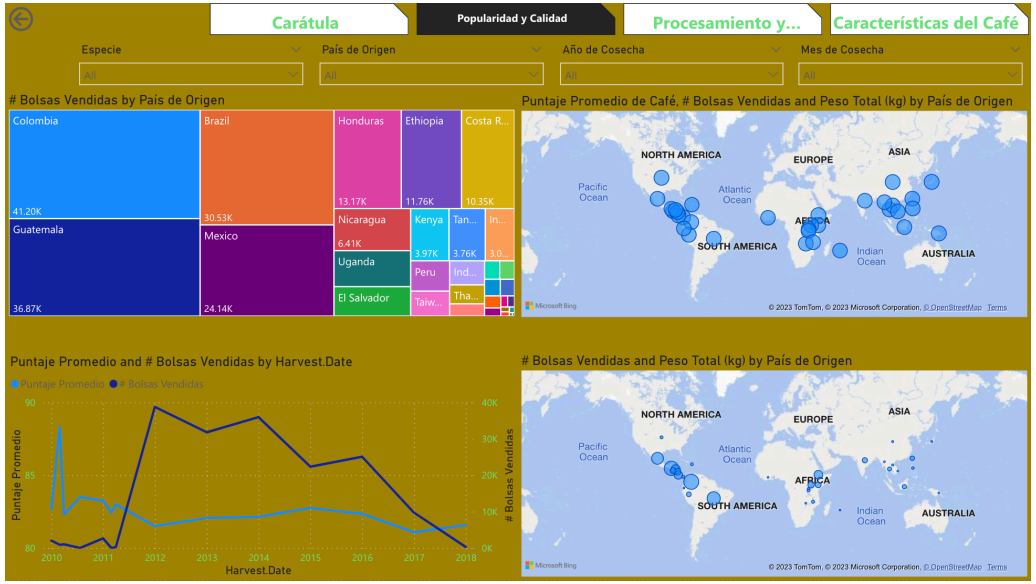
ENTREGA DEL PROYECTO FINAL

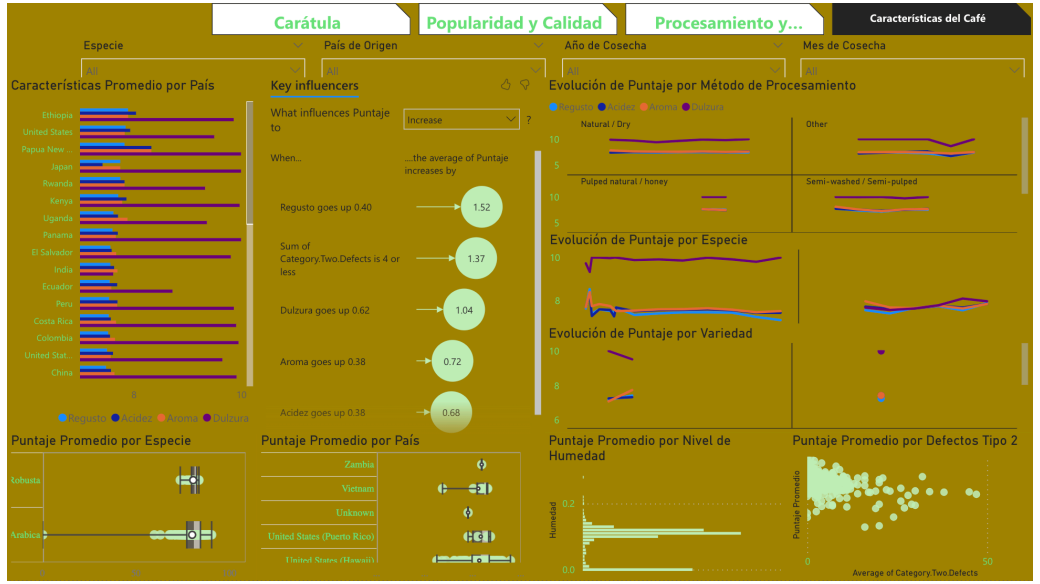
Visualización de los datos

11. Listado de páginas del reporte

El reporte en PowerBI está conformado por 4 páginas. La primera corresponde a una carátula que introduce la temática del reporte y contiene pestañas de navegación hacia el resto del reporte permitiendo una navegación lineal o a elección del usuario según el foco de interés.

Las otras páginas consisten en conjuntos de visualizaciones organizadas temáticamente cada una con un objetivo particular de presentación y que en conjunto soportan el objetivo del reporte que es proporcionar un vistazo de lo más general y descriptivo a lo más especifico y explicativo de la calidad del café y los factores que impactan en la calidad de la taza de café.





ENTREGA DEL PROYECTO FINAL

Visualización de los datos

12. Conclusiones

En primer lugar podemos identificar que la cosecha de café requiere una combinación particular de condiciones ambientales en general ya que vemos que la producción de cafê se centra en zonas tropicales y en la banda central latitudinal. Sin embargo podemos ver que la popularidad del café tiene una impronta importante de costumbre y branding país ya que aunque notamos que hay ciertos países que concentran la gran mayoría de la venta de bolsas de café como Colombia, Guatemala, Brazil y México y luego la participación de los otros países decae rápidamente, también notamos que la calidad de la taza de café según las clasificadoras internacionales es bastante uniforme moviéndose entre 80-84 de puntaje para la gran mayoría de cafes indistintamente de su origen. Esto también podría ser indicativo que los países mas populares producen variedades que son perfiles mas populares en el consumo masivo.

Respecto a los métodos de procesamiento y variedades de café podemos notar que el método de procesamiento de lavado en húmedo (Washed/wet) es el método mas popular y el método con mayor puntaje en general. Esto puede deberse al que es un proceso más estandarizado y tecnificado. Por el lado de las variedades vemos que la Caturra y Bourbon son las variedades de mayor puntaje y popularidad utilizando los filtros podemos notar que son ambas variedades del café Arabica.

Una vez que hacemos zoom en las características del café en si podemos ver que la característica del café que mas se ha desarrollado en todos los países es la dulzura que tiene mayor puntaje consistentemente en todos los orígenes, seguido por el Aroma, la Acidez y el Regusto, en ese orden. No obstante cuando observamos la visualización de influenciados clave del puntaje del café vemos que el regusto del café es el factor con mayor impacto en el puntaje, seguido de la frecuencia de defectos tipo 2 (restos de ramas, fruto y otras materias en el café), la dulzura le siguen en ese orden. Esto podría ser util para los productores de café ya que indica que podrían estar priorizando la dulzura cuando si se enfocaran mas en el regusto podrían generar un mayor impacto positivo en su calificación oficial. No obstante también cabe considerar que es posible que en el mercado de consumo masivo características como la dulzura pueden ser un driver de la preferencia de la persona común por lo que esta decisión puede ser acertada ya que si se perjudica el atractivo comercial del café para obtener un mejor puntaje, que es de por si un indicador de calidad usado por el mercado especialista o aficionado, podría perjudicar la rentabilidad de las cafetaleras. Esta conclusión se ve soportada por el hecho que identificamos anteriormente que la calidad de los cafés es en general alta y no varia mucho lo cual podría indicar que tratar de mejorar aun más el puntaje podría generar rendimientos decrecientes en las ventas.

Tambien podemos identificar que aunque los puntajes son generalmente altos para todos los países, metidos de procesamiento, especie, variedad, etc. Lo que si se puede identificar es que la consistencia es importante para la calidad de la taza del café y vemos que mientras mas defectos el puntaje tiende a ser menor y complementariamente que las variedades y especies y métodos mas comerciales y populares también tienden a tener una mayor variabilidad en el puntaje. Esto seria otro hecho que soporta la conclusion que el mercado del café esta determinado por un trade-off entre calidad de café y popularidad.