|  |
| --- |
| el  café  de especialidad |
| análisis de datos exploratorios |
|  |
|  |

# cono de paola | the bridge school | data science

indice

[INTRODUCCIÓN 1](#_Toc141998737)

[¿qué es el café? 1](#_Toc141998738)

[¿qué es el café de especialidad? 1](#_Toc141998739)

[análisis exploratorio de datos 1](#_Toc141998740)

[Hipótesis: 1](#_Toc141998741)

[Obtención de los Datos: 1](#_Toc141998742)

[Primer Data Frame: 1](#_Toc141998743)

[Eliminación de Columnas: 1](#_Toc141998744)

[Renombrado de Columnas: 1](#_Toc141998745)

[Reordenamiento de Columnas: 1](#_Toc141998746)

[Ajuste del Total de kg por Saco: 2](#_Toc141998747)

[Cálculo del Total de Café de Especialidad: 2](#_Toc141998748)

[Agrupación y Cálculo de la Media de Puntuación por País: 2](#_Toc141998749)

[Gráfico de Puntuación Media por País de Café de Especialidad: 2](#_Toc141998750)

[Cálculo de Correlaciones: 3](#_Toc141998751)

[Gráfico de Correlación: 3](#_Toc141998752)

[Segundo Data Frame: 3](#_Toc141998753)

[Eliminación de Años no Utilizados: 4](#_Toc141998754)

[Análisis de la Producción de Café de Especialidad: 5](#_Toc141998755)

[Creación de un Dataframe Específico: 5](#_Toc141998756)

[Integración de Información de Producción Total de Café: 6](#_Toc141998757)

[Eliminación de Países sin Datos de Producción: 6](#_Toc141998758)

[Test de Hipótesis para Identificar Relaciones: 6](#_Toc141998759)

[Conclusión test de hipótesis: 8](#_Toc141998760)

[Análisis de los tres países relevantes en el mercado mundial de café por producción y calidad: 8](#_Toc141998761)

[Brasil: 9](#_Toc141998762)

[Colombia: 11](#_Toc141998763)

[Etiopía: 13](#_Toc141998764)

[Imputaciones: 14](#_Toc141998765)

# INTRODUCCIÓN

En este informe se presenta el análisis exploratorio de datos (EDA) realizado sobre un conjunto de datos relacionado con la producción y calidad de café de especialidad en distintos países. El objetivo del proyecto fue explorar la relación entre diversas variables y descubrir patrones, tendencias o correlaciones que pudieran proporcionar insights significativos.

# ¿qué es el café?

El café es una bebida aromática y estimulante que se obtiene a partir de las semillas tostadas y molidas de la planta del café, conocida científicamente como Coffea. Es una de las bebidas más populares en el mundo y se consume en diferentes formas y preparaciones en diversas culturas.

Las principales especies de café que se utilizan para la producción comercial son Arábica y Robusta. El café arábico es el más ampliamente consumido debido a su sabor más suave y complejo, mientras que el café robusto es valorado por su mayor contenido de cafeína y su contribución a la creación de cafés más fuertes y con cuerpo.

# ¿qué es el café de especialidad?

El café de especialidad es un café de alta calidad que se cultiva, procesa y prepara de forma cuidadosa y meticulosa para resaltar sus sabores y características únicas. Su principal rasgo, es su gran diferenciación con los cafés comerciales y su capacidad de volver el consumo de café en una experiencia sensorial más compleja. Para que un café sea considerado de especialidad es necesario:

* Ser de variedad arábica.
* Obtener una puntuación mínima de 80 puntos sobre 100 de acuerdo con el sistema internacional para evaluación de calidad, establecido por la Specialty Coffee Assotation (SCA).
  + Los atributos que evalúan los catadores profesionales (Q-Graders) son:
* Acidez.
* Aroma.
* Sabor.
* Cuerpo
* Post-gusto
* Balance
* Uniformidad
* Dulzor.

# análisis exploratorio de datos

## Hipótesis:

1. A mayor producción de café en un país, es más probable que se observe una mayor cantidad de café de especialidad en el mismo. La hipótesis se basa en la suposición de que a medida que la producción total de café aumenta en un país, también aumenta la posibilidad de que se cultiven variedades de café de mayor calidad y se produzca café de especialidad.
2. El país que es el mayor productor de café a nivel mundial también es el que más café de especialidad aporta en comparación con otros países productores. Esta hipótesis se basa en la idea de que el país que lidera la producción mundial de café, además de cultivar y exportar grandes volúmenes de café convencional, también tiene la capacidad y los recursos para producir café de especialidad en mayor cantidad en comparación con otros países productores.

## Obtención de los Datos:

Para la elaboración del EDA, hemos seleccionado un data set compuesto de 18 variables y 985 registros de lotes de café de especialidad de un total de 32 países desde el año 2010 hasta el año 2018. Adicionalmente, obtuvimos otra base de datos con la producción mundial de café por país desde 1990 hasta 2019, información complementaria que nos ayudará a la comprobación de las hipótesis. El primer data Set es originario del Coffee Quality Institute (<https://database.coffeeinstitute.org/>), el segundo, es originario del International Coffee Organization (<https://www.ico.org/>)

## Primer Data Frame:

Café de Especialidad (df\_quality)

## Eliminación de Columnas:

Con el fin de simplificar el conjunto de datos y centrarnos en las variables más relevantes para el análisis, se eliminaron las columnas "Location.Region", "Location.Altitude.Min", "Location.Altitude.Max" y "Location.Altitude.Average". Estas columnas se consideraron menos relevantes para el enfoque del estudio ya que la mayoría de los valores eran nulos.

## Renombrado de Columnas:

Para mejorar la claridad y legibilidad del conjunto de datos, todas las columnas se renombraron utilizando un diccionario. Esto permitió que las columnas tuvieran nombres más descriptivos y fáciles de entender.

## Reordenamiento de Columnas:

Con el propósito de agrupar las variables descriptivas y cualitativas, se reordenaron las columnas para mejorar la organización del conjunto de datos.

## Ajuste del Total de kg por Saco:

Para tener en cuenta posibles valores atípicos o extremos en el conjunto de datos, se ajustó el total de kg por saco calculando el percentil 50 (mediana). Esto permitió obtener un valor más representativo de la cantidad de café en cada saco.

## Cálculo del Total de Café de Especialidad:

Mediante la multiplicación del número de sacos por su peso ajustado, se obtuvo el valor total del café de especialidad para cada lote, que se añadió a una nueva columna llamada "Total\_Specialty\_coffee". Esta columna representó la cantidad total de café de especialidad en cada lote.

## Agrupación y Cálculo de la Media de Puntuación por País:

Los lotes de café se agruparon por país y se calculó la media de puntuación de cada atributo de calidad, como "Score\_Aroma", "Score\_Flavor", "Score\_Aftertaste", "Score\_Acidity", "Score\_Body", "Score\_Balance", "Score\_Uniformity" y "Score\_Sweetness". Esto permitió obtener una visión general de la calidad del café de especialidad producido en cada país.

## Gráfico de Puntuación Media por País de Café de Especialidad:

Para visualizar y comparar la calidad del café de especialidad por país, se creó un gráfico que muestra la puntuación media para cada atributo de calidad en el eje vertical, y los países en el eje horizontal. Este gráfico proporcionó una representación visual de las diferencias en la calidad del café entre los países analizados, hemos resaltado en color verde aquellos que se encontraban por encima de la media (Etiopía y Papúa Nueva Guinea) y en color rojo los que estaban por debajo de los 80 puntos mínimos para ser considerados cafés especiales (Nicaragua, Honduras, Haití)

Gráfico, Gráfico de burbujas

Descripción generada automáticamente

## Cálculo de Correlaciones:

Se realizó un análisis de correlación para examinar las relaciones entre diversas variables. Se calculó la correlación entre las variables "Country", "Production\_Bag", "KG\_x\_Bag", "Total\_Specialty\_coffee" y las diferentes puntuaciones de calidad ("Score\_Aroma", "Score\_Flavor", "Score\_Aftertaste", "Score\_Acidity", "Score\_Body", "Score\_Balance", "Score\_Uniformity" y "Score\_Sweetness").

## Gráfico de Correlación:

Con el propósito de visualizar las correlaciones entre las variables, se creó un gráfico de correlación que ayudó a identificar posibles relaciones positivas o negativas entre ellas.

Observando el gráfico de correlación, podemos observar que existe correlación entra las características cualitativas del café, sin embargo, esta información no resulta del todo relevante, ya que siendo una base de datos de cafés de especialidad se espera que los valores no se diferencien mucho entre ellos.

Imagen que contiene Escala de tiempo

Descripción generada automáticamente

## Segundo Data Frame:

En el siguiente Gráfico de mapa agrupamos los países productores por continente para poder observarlos de forma más sencillas. Adicionalmente, se les ha asignado el tamaño de la burbuja en función de la producción total de café de cada uno.

Mapa

Descripción generada automáticamente

## Eliminación de Años no Utilizados:

Dentro del proceso de preparación de los datos, se realizó una cuidadosa selección de los años de interés para este análisis. Aquellos años que no aportaban información relevante o que no eran representativos del período de estudio fueron eliminados. De esta manera, se logró enfocar el análisis en los datos más pertinentes para abordar la relación entre la producción de café de especialidad y la producción total de café por país.

**Gráfico, Histograma

Descripción generada automáticamenteDensidad de los datos por años de producción**

## Análisis de la Producción de Café de Especialidad:

Con el conjunto de datos preparado, se procedió a realizar un análisis específico para obtener el total de producción de café de especialidad agrupado por país y año. Este análisis permitió obtener una perspectiva general de la evolución de la producción de café de especialidad a lo largo del tiempo y su distribución geográfica. La agrupación por país y año proporcionó métricas resumidas y valiosas para comprender las tendencias y patrones relacionados con la producción de café de especialidad.

**Producción de café por año** **Gráfico, Gráfico de líneas

Descripción generada automáticamente**

**Producción de café de especialidad por año**

**Gráfico, Gráfico de líneas

Descripción generada automáticamente**

## Creación de un Dataframe Específico:

A partir de los resultados del análisis de la producción de café de especialidad, se generó un nuevo dataframe específico para el estudio (production\_world\_quality). Este nuevo dataframe contenía únicamente los datos relevantes sobre la producción de café de especialidad por país y año. La creación de este dataframe permitió centrar el análisis en la variable de interés y proporcionó una base sólida para investigar su relación con la producción total de café por país.

## Integración de Información de Producción Total de Café:

Para analizar la relación entre la producción de café de especialidad y la producción total de café por país, se integró la información del dataframe de estudio con los datos de producción total de café por país y año. Esta integración de datos permitió establecer conexiones y realizar comparaciones entre ambas variables en cada país y período de tiempo específico.

**Producción Total de Café**

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza baja

## Eliminación de Países sin Datos de Producción:

En aras de mantener la coherencia y la fiabilidad de los resultados, se llevó a cabo la eliminación de aquellos países para los cuales no se disponía de datos de producción total por año. Esta medida aseguró que el análisis se centrara únicamente en los países con información completa y evitó posibles sesgos en los resultados de las pruebas de hipótesis.

## Test de Hipótesis para Identificar Relaciones:

Finalmente, para determinar si existe una relación significativa entre la producción de café de especialidad y la producción total de café por país, se realizaron tres (3) test de hipótesis no paramétricos, debido a que la producción total de café y la de café de especialidad, no siguen una distribución normal, la cual se comprobó con un QQ-plot:

**QQ-plot Producción total de café y café de especialidad**

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Prueba de Wilcoxon:

* 1. La Prueba de Wilcoxon es una prueba no paramétrica que se utiliza para comparar dos muestras relacionadas o pareadas. En este contexto, se aplicó la prueba de Wilcoxon para analizar si había una relación significativa entre la producción de café de especialidad y la producción total de café en cada país a lo largo del tiempo. Al ser una prueba no paramétrica, es adecuada cuando los datos no siguen una distribución normal. Además, esta prueba es útil cuando no se puede asumir una relación lineal entre las variables.
  2. Resultado: WilcoxonResult(statistic=340.0, pvalue=1.5552518626156703e-16)

Coeficiente de correlación de Spearman:

* 1. El Coeficiente de correlación de Spearman es otra medida no paramétrica utilizada para evaluar la relación entre dos variables. Se aplicó esta prueba para determinar si existía una correlación monotónica entre la producción de café de especialidad y la producción total de café en cada país y año específico. La correlación de Spearman es adecuada para datos ordinales o no paramétricos y es menos sensible a valores atípicos.
  2. Resultado: SignificanceResult (statistic = 0.2775817297525635, pvalue = 0.002914131539632695)

Coeficiente de correlación de Kendall tau:

* 1. El Coeficiente de correlación de Kendall tau también es una medida no paramétrica que se utiliza para evaluar la asociación entre dos variables ordinales. Se aplicó esta prueba para analizar si había una correlación significativa entre la producción de café de especialidad y la producción total de café en cada país durante el período de estudio. Al igual que la correlación de Spearman, el coeficiente de correlación de Kendall tau es adecuado para datos no paramétricos y es menos sensible a valores atípicos.
  2. Resultado: SignificanceResult(statistic=0.18714409964749032, pvalue = 0.003354636819726005)

## Conclusión test de hipótesis:

Según los resultados obtenidos a través de los test de Wilcoxon, Spearman y Kendall Tau, debemos rechazar la hipótesis planteada. Los resultados de estos tres tests estadísticos indican que no existe una correlación significativa entre la producción total de café en un país y la cantidad de café de especialidad observado en el mismo.

La hipótesis inicial sugiere que a medida que la producción total de café aumenta en un país, también se incrementa la probabilidad de que se cultiven variedades de café de mayor calidad y se produzca café de especialidad. Sin embargo, los resultados de los tests estadísticos muestran que no hay evidencia suficiente para respaldar esta afirmación.

# Análisis de los tres países relevantes en el mercado mundial de café por producción y calidad:

1. Brasil
2. Colombia
3. Etiopía

Producción Total de Café VS Producción de Café de Especialidad

Gráfico

Descripción generada automáticamente

Como podemos observar, no existe evidencia suficiente que nos indique que el país mayor productor de café también produzca la mayor cantidad de cafés especiales. Es por ello que analizaremos por separado el top 5 de países productores de café.

## Brasil:

Producción Total de Café por año de Brasil VS Café de Especialidad

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Evolución Calidad del café de Especialidad en Brasil

Gráfico, Gráfico de líneas

Descripción generada automáticamente

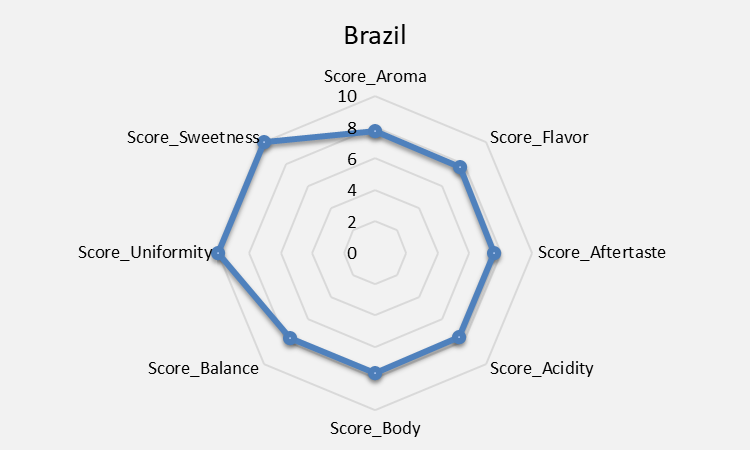
Luego del Análisis Exploratorio de Datos (EDA) sobre el café de Brasil, hemos identificado hallazgos significativos que proporcionan una visión detallada de la industria cafetera del país. En primer lugar, encontramos que la producción total de café en Brasil no muestra una relación significativa con la producción de cafés especiales. Esto indica que, aunque Brasil es un importante productor de café a nivel mundial, la producción de cafés especiales es independiente de su producción total.

Sin embargo, uno de los puntos más alentadores de nuestro análisis es que se observó un aumento en la producción de café de especialidad entre los años 2010 y 2016. Esta tendencia positiva sugiere que el enfoque en la producción de cafés especiales está creciendo, lo que puede tener un impacto significativo en el posicionamiento y prestigio del café brasileño en el mercado internacional.

Por otro lado, se detectó una disminución en las puntuaciones de calidad del café brasileño entre 2010 y 2016. Este hallazgo es de gran relevancia, ya que destaca la necesidad de implementar medidas adecuadas para mantener y mejorar los estándares de calidad en la producción de café en Brasil.

En resumen, el EDA revela que la producción de café de especialidad en Brasil experimentó un aumento durante el periodo analizado, mientras que las puntuaciones de calidad disminuyeron. Estos hallazgos presentan una oportunidad para fortalecer la calidad y la reputación del café brasileño en el mercado global y permiten enfocar los esfuerzos en mejorar la producción de cafés especiales en el futuro.

Puntuación Promedio de los Atributos del Café de Brasil



## Colombia:

Producción Total de Café por año de Colombia VS Café de Especialidad

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Evolución Calidad del café de Especialidad en Colombia

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Después de realizar el Análisis Exploratorio de Datos (EDA) sobre el café de Colombia, hemos obtenido conclusiones significativas que arrojan luz sobre la industria cafetera del país. A continuación, se presentan los hallazgos más importantes:

Se ha identificado una clara tendencia de crecimiento en la producción total de café en Colombia durante el período comprendido entre 2011 y 2017. Este crecimiento constante en la producción sugiere un aumento en la productividad y en la capacidad del país para satisfacer la demanda tanto nacional como internacional de café colombiano.

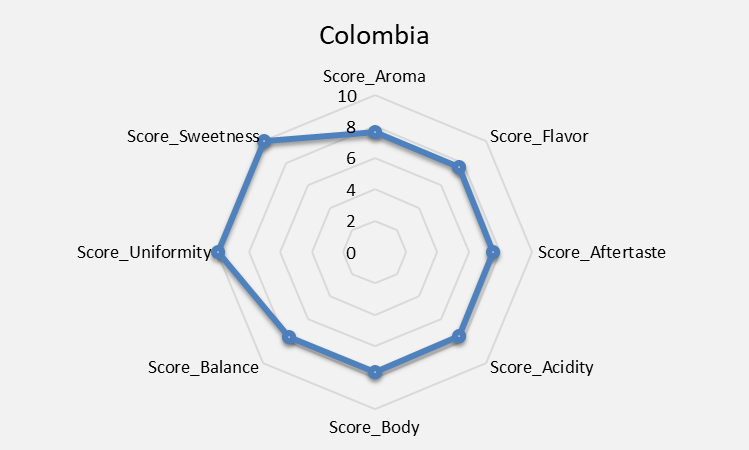
En cuanto al café de especialidad, se ha observado una disminución significativa en su producción entre 2011 y 2017. Esto puede tener implicaciones en la oferta y la calidad de los cafés especiales colombianos, lo que destaca la necesidad de analizar las causas detrás de esta disminución y buscar oportunidades para mejorar y promover la producción de cafés especiales.

Es importante destacar que, a pesar de la disminución en la producción de cafés especiales, la puntuación promedio del café de especialidad ha incrementado entre 2011 y 2017. Esta mejora en las puntuaciones sugiere que, aunque la producción puede haber disminuido, se ha puesto un mayor énfasis en la calidad y el cuidado en la producción de los cafés especiales en Colombia.

Aunque la producción total de café en Colombia ha experimentado un crecimiento, no se ha encontrado una relación significativa entre la producción total y la producción de cafés especiales. Esto indica que ambos tipos de café están influenciados por factores diferentes, como las condiciones climáticas, la tecnología de producción y las políticas agrícolas.

En resumen, el EDA sobre el café de Colombia revela una tendencia positiva en la producción total de café, pero una disminución en la producción de cafés especiales. Sin embargo, se ha notado un incremento en la calidad promedio de los cafés especiales producidos.

Puntuación Promedio de los Atributos del Café de Colombia



## Etiopía:

Producción Total de Café por año de Etiopía VS Café de Especialidad

**Gráfico, Gráfico de líneas

Descripción generada automáticamente**

Evolución Calidad del café de Especialidad en Etiopía

**Gráfico, Gráfico de líneas

Descripción generada automáticamente**

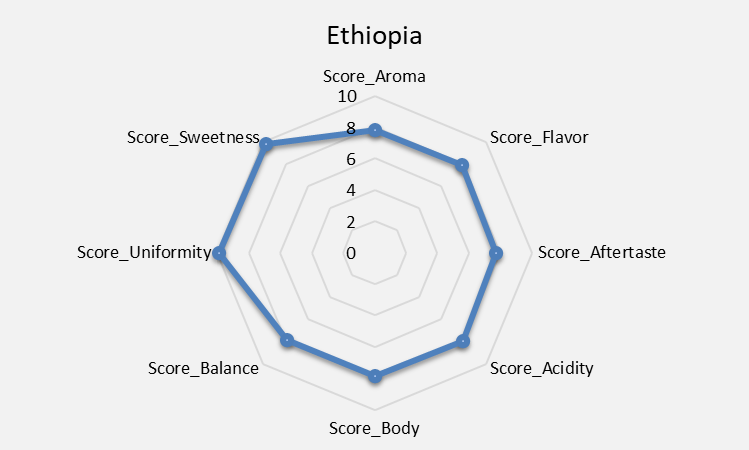
La producción de café en Etiopía muestra una tendencia de incremento cada año. Este crecimiento constante indica una mejora en la productividad y en la capacidad del país para satisfacer la demanda tanto nacional como internacional de café etíope.

Se ha encontrado una leve relación entre la producción de café especiales y la producción total de café en Etiopía. Aunque no es una correlación fuerte, esto sugiere que, en cierta medida, el aumento en la producción de cafés especiales puede estar influenciado por el crecimiento general de la producción total.

El año 2014 fue el de mayor producción de cafés especiales en Etiopía y registró la máxima puntuación de todos los países, con 90 puntos. Este hito destaca la calidad excepcional de los cafés especiales producidos en Etiopía durante ese año, lo que puede haber contribuido a la consolidación de su posición en el mercado internacional.

Durante el período comprendido entre 2010 y 2017, se ha observado un leve incremento en la producción de cafés especiales en Etiopía. Este aumento sugiere un mayor enfoque en la producción y promoción de cafés especiales, lo que puede ser una estrategia efectiva para agregar valor y mejorar la posición del café etíope en el mercado.

Puntuación Promedio de los Atributos del Café de Colombia



# Imputaciones:

Durante el estudio de la segunda hipótesis, hemos dividido el total de café de especialidad producido por país y por año, entre la producción total de café por año. Este ejercicio, nos arrojó que debíamos imputar una serie de países que no proporcionaban información relevante al EDA. Los países fueron: China, Honduras, Guatemala, Estados Unidos y Myanmar.

Un ejemplo de ello, lo encontramos en el siguiente gráfico, donde podemos observar que la producción de café de especialidad en Guatemala es mayor a la producción total.

Gráfico, Gráfico de líneas

Descripción generada automáticamente