**Entrega 1:** Ciencia de Datos

Valeria Herrera Ríos

1192721349

**Resumen:**

El café es la segunda bebida más consumida en el mundo después del agua, en la actualidad tanto la cultura del café como las tendencias alrededor de él han evolucionado exponencialmente, haciendo que el mundo del café continué fortaleciéndose e innovando, priorizando ahora la calidad sobre el precio.

El dataset coffee\_analysis2 contiene información detallada sobre diversos perfiles de café, incluyendo el nombre del café, el tostador, el nivel de tostión, el país, el origen, el precio, el puntaje sensorial, la fecha de evaluación y descripciones cualitativas. Este conjunto de datos permite explorar la influencia de factores como el nivel de tostión y el origen geográfico en la puntuación obtenida en catas sensoriales, y como estas variables influyen en el precio final del café.

A través de visualizaciones univariadas, bivariadas y multivariadas, se busca identificar patrones que revelen cómo aspectos como el país de origen o el nivel de tostión se relacionan con la calidad y el valor del producto. Además, se analizará la evolución temporal de las evaluaciones para detectar tendencias estacionales y su posible impacto en los precios.

El objetivo central es responder preguntas e hipótesis relacionadas con la calidad sensorial y la valorización económica del café, proporcionando interpretaciones basadas en datos que puedan apoyar decisiones en procesos de producción y comercialización. Es importante resaltar que el presente dataset contiene información muy valiosa, sin embargo, no contiene todos los datos detallados sobre el perfil del café, haciendo que su análisis se reduzca a ciertos aspectos como el origen geográfico y la tostión, pero no se cuenta con información relevante como el proceso de beneficio y la variedad del café.

**Hipótesis**

* **Hipótesis 1:** ¿Cómo afecta el nivel de tostión al puntaje sensorial y al precio?

Los cafés con tostión media presentan puntajes sensoriales más altos en comparación con tostiones altas y bajas.

* **Hipótesis 2:** ¿Existe relación entre el país de origen, el puntaje y el precio?

Cafés de países reconocidos por alta calidad como Colombia, Kenya, Panama, Ethiopia... tienden a obtener puntajes y precios superiores.

* **Hipótesis 3:** ¿El precio está directamente relacionado con el puntaje?

Un café con mayor puntaje de catación tiende a tener un precio más alto

**Proceso:**

1. Inicialmente se conectó la herramienta Colab (Google Colaboratory) con drive, para posteriormente llamar el dataset ubicado en una carpeta de drive llamada Coder, este paso de llamar el dataset fue complejo ya que el código que conocía inicialmente no funciono, con ayuda la IA logre hacer la conexión.

Imagen 1.



1. Posteriormente se importó la librería pandas.

Imagen 2.



1. Se cargo el archivo en un dataframe y se determinaron las columnas requeridas, sin embargo, presento un inconveniente, pues la selección y ubicación de las columnas se hizo de forma desordenada, por esta razón fue necesario plantearlo de forma diferente, con los separadores adecuados

Imagen 3.



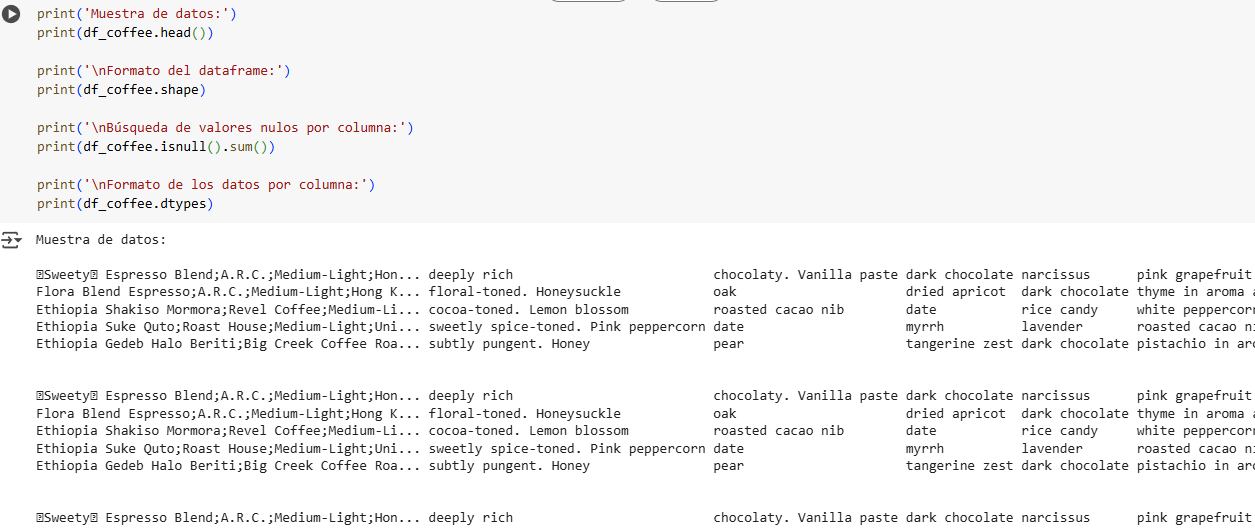
Es por esto que se probó con diferentes separadores, hasta lograr determinar que el correcto es sep: “;”

Imagen 4.



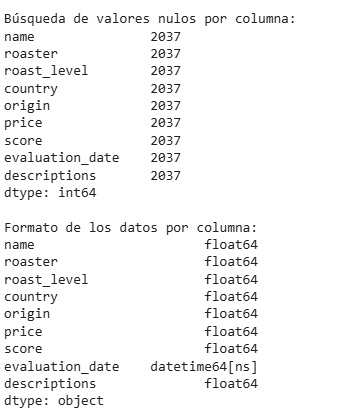
1. Se realizó una evaluación de los datos que se encuentran en el dataframe

Imagen 5.



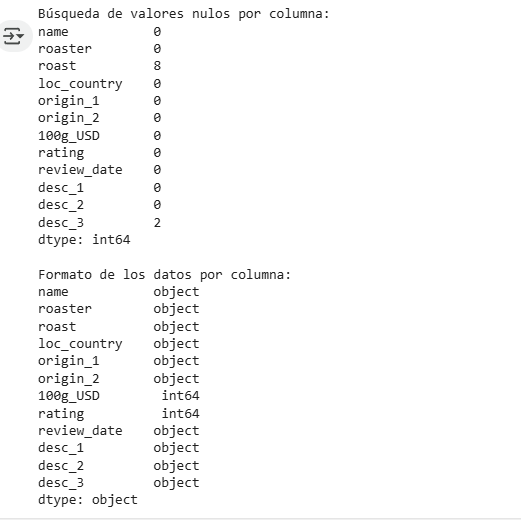
En este apartado se presento un inconveniente porque pandas reconoció todos los datos del dataframe como nulos, lo que hizo que los datos pasaran de ser strings a float tal como se denominan los datos nulos y como se observa en la imagen 6.

Imagen 6



Después de descubrir que eran los separadores ahora el análisis arroja los datos correctos:

Imagen 7

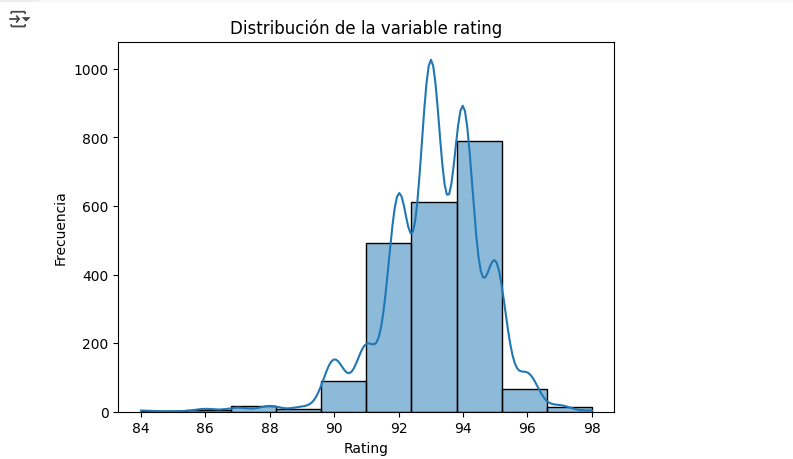


**Visualizaciones Univariadas:**

1. La visualización de tipo histograma permite observar que la mayor concentración de datos se encuentra en el rango de puntajes entre 91 y 95, siendo el rating de 94 el más frecuente, con aproximadamente 800 registros. Esta distribución indica que una gran parte de los cafés analizados poseen calificaciones excepcionales, lo cual concuerda con los estándares de la Specialty Coffee Association (SCA.

La alta frecuencia de puntajes en este rango sugiere que la muestra de cafés evaluados se encuentra compuesta predominantemente por cafésunida calidad sensorial alta. Esto podría deberse a un sesgo en los datos recolectados, como las evaluaciones enfocadas en cafés de especialidad (puntajes mayores de 80 según la SCA) o a una representación deliberada de productos premium en el dataset.

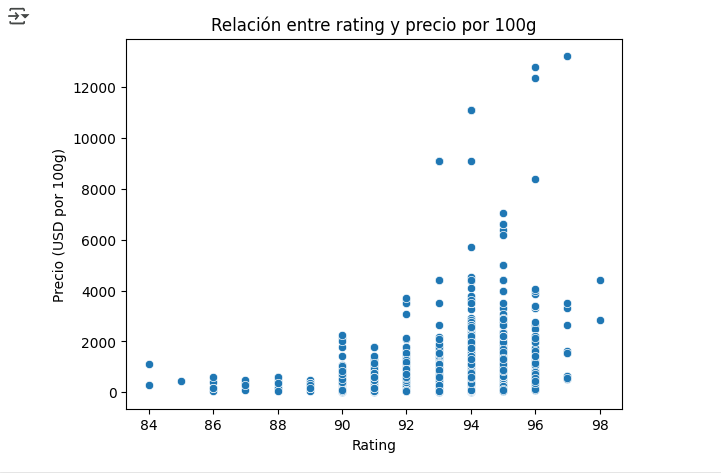
Gráfico 1. Histograma



**Visualizaciones bivariadas:**

1. La visualización muestra una tendencia general ascendente entre el rating del café y su precio con respecto a cada 100g de café, lo que sugiere que los cafés mejor calificados suelen tener precios más altos. Sin embargo, esta relación no es perfectamente lineal, ya que se evidencia una alta dispersión en los datos, especialmente entre ratings de 90 a 95.

Gráfico 2: Grafico de dispersión



1. Esta visualización permite comparar cómo varía la calidad percibida del café, representada en puntuaciones entre diferentes países. A través de la mediana, el rango intercuartílico y la presencia de valores atípicos, se pueden identificar patrones y posibles diferencias.

Puntajes consistentemente altos (mediana ≈ 94), tanto Hong Kong, Canadá, Taiwán, Japón, y Hawai presentan medianas cercanas o iguales a 94, lo que indica una evaluación positiva constante. Estos países tienen una distribución de puntajes sugiere cafés de alta calidad, con menor presencia de puntajes bajos.

Se observa que se presenta una variabilidad amplia en EE. UU, aunque su mediana es alta, Estados Unidos muestra una gran cantidad de valores atípicos hacia la baja, lo que sugiere una mayor variabilidad en la calidad de los cafés evaluados.

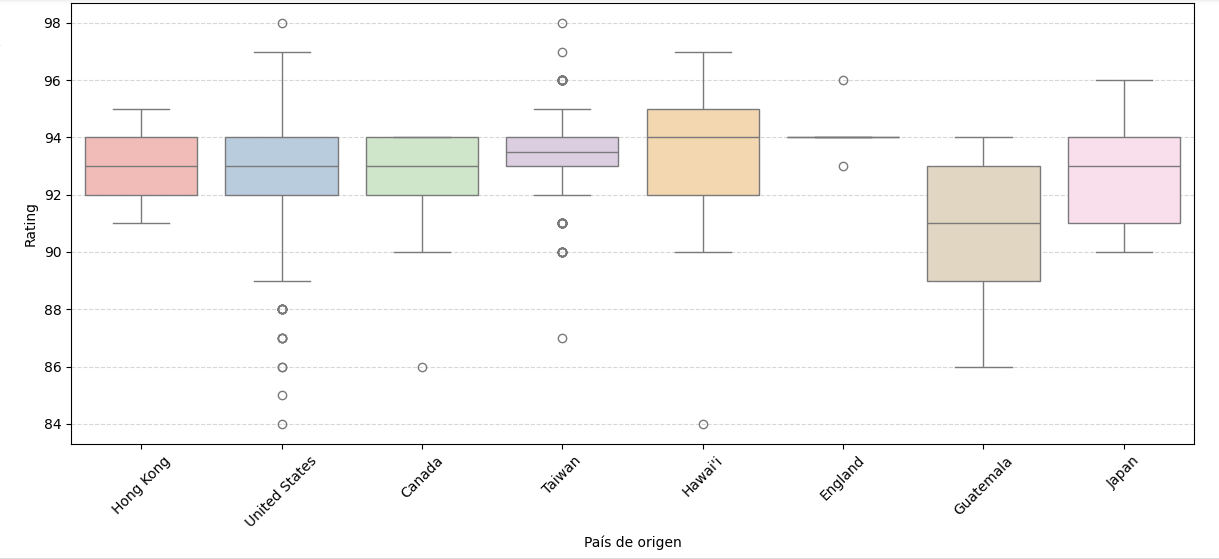
Por el contrario, Guatemala presenta menor mediana, está por debajo de la mayoría, y su rango intercuartílico es más bajo, lo cual podría indicar una calidad percibida ligeramente inferior respecto a otros países en el conjunto analizado. Además, tiene una dispersión más amplia, lo que indica inconsistencia en las calificaciones.

Varios países contienen inconsistencias y valores extremos, presentan valores extremos hacía abajo, especialmente EE.UU., Taiwán y Hawai. Esto puede deberse a cafés defectuosos o con las cualidades en taza no esperados.

Por último, Inglaterra sin duda es un caso diferente, presenta valores constantes; no tiene rango intercuartílico visible, lo cual indica que los puntajes fueron idénticos o muy similares para todos los cafés evaluados. Esto se debe a una muy baja muestra.

Gráfico 3. Boxplot.

Distribución del rating por país de origen



**Visualización multivariada**

1. La visualización muestra cómo se distribuyen las frecuencias de distintos tipos de tueste según el país de origen del café. Esto permite identificar patrones geográficos en las preferencias o calidad según el tueste.

Estados Unidos es el país con la mayor cantidad de registros en el dataset, destacando especialmente en el tueste Medium-Light, con 995 registros. También tiene alta representación en los tuestes Light (166) y Medium (151), lo que sugiere una preferencia por tuestes más suaves o claros.

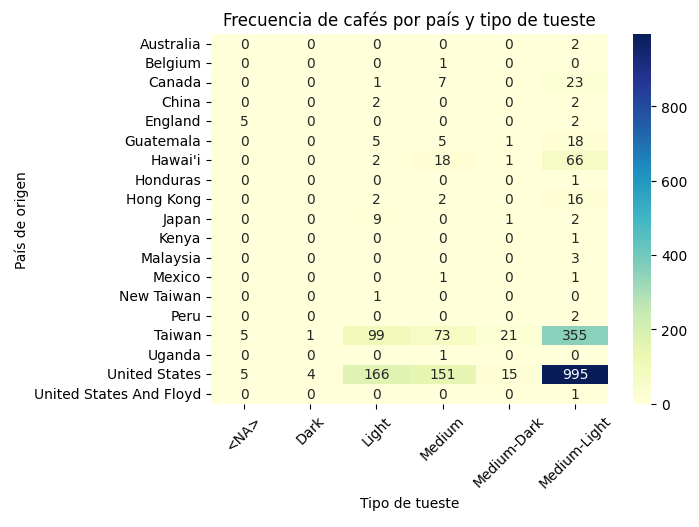
Taiwan también presenta una distribución relevante, especialmente en Medium-Light (355), seguido de Light (99) y Medium (73). Esto indica una afinidad similar a la de EE.UU. por tuestes más claros.

Aunque con menor volumen, Hawai y Japón se inclinan también por tuestes Medium-Light y Medium, mostrando una posible tendencia compartida hacia sabores más brillantes o frutales.

Muchos países como Australia, Honduras, Kenya o Perú tienen muy pocos registros, lo que limita el análisis para esas regiones y podría indicar un sesgo en la base de datos hacia ciertos países exportadores o tostadores.

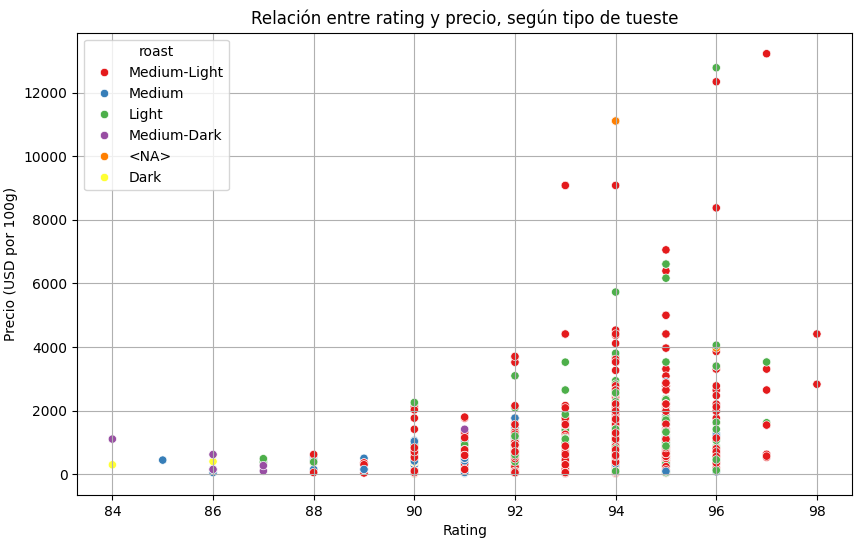
Los valores nulos o NA, es un error en la lectura del dataset, pues el mismo no presenta datos nulos o inexistentes.

Grafica 4. Mapa de calor



1. Se observa una clara tendencia positiva entre el puntaje y el precio: a mayor rating, el precio tiende a ser más alto. Respecto al tipo de tostión, las tostiones "Medium-Light", "Medium" y "Light" son las que consistentemente alcanzan los ratings y precios más elevados, lo que sugiere una asociación positiva entre estas tostiones y cafés de mayor calidad y costo. En contraste, la tostión "Dark" muestra una tendencia inversa, asociándose con los ratings y precios más bajos. No hay evidencia contundente de que solo las tostiones medias sean las que obtengan los puntajes más altos; más bien, el rango de tostiones media clara es el que se correlaciona con la excelencia.

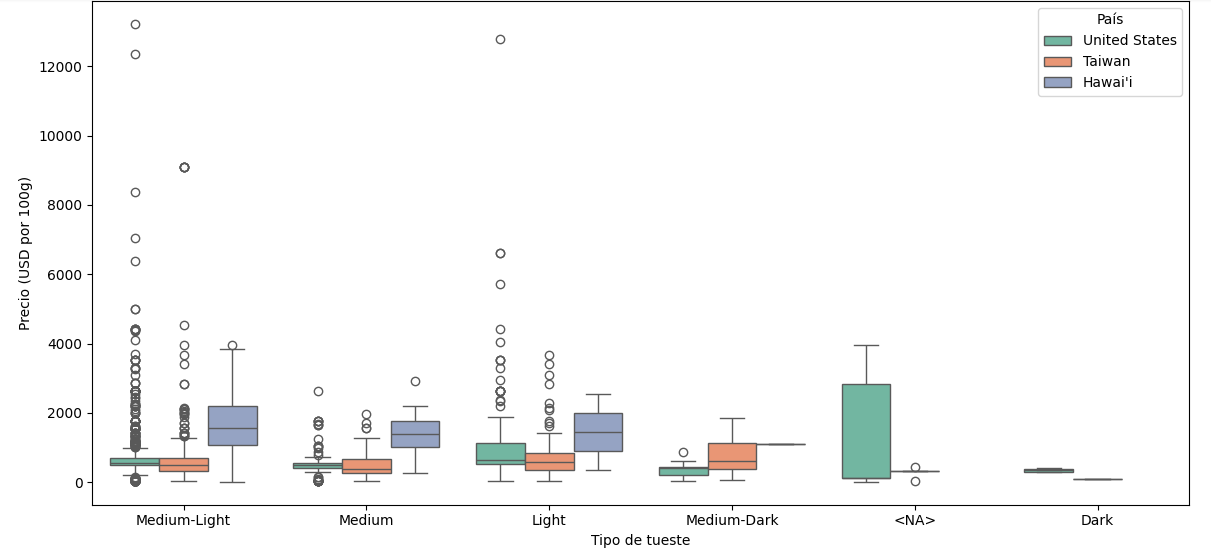
Gráfico 5. Gráfico de dispersión



1. Este diagrama de cajas multivariado muestra que el precio de un café está fuertemente influenciado tanto por el tipo de tueste como por el país de origen. Existe una clara tendencia de precios más altos para tuestes "Medium-Light", "Medium" y "Light", que también exhiben una mayor variabilidad. Sin embargo, la influencia del país de origen es dominante: los cafés de Hawái demuestran una tendencia abrumadora hacia precios significativamente más elevados en comparación con Taiwán y Estados Unidos, independientemente del tipo de tueste. Los cafés de Taiwán generalmente superan en precio a los de Estados Unidos. Esto sugiere que el origen geográfico, más allá del tueste, es un fuerte predictor del precio.

Gráfico 6. Boxplot

Precio por tipo de tueste y país de origen

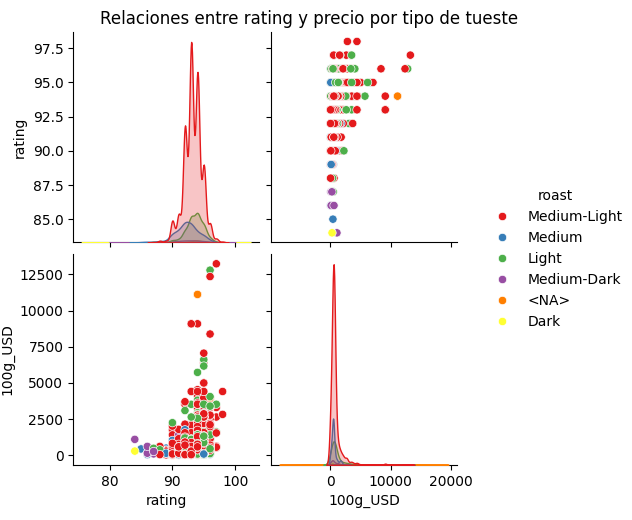


1. Se revela una fuerte tendencia positiva entre el puntaje de catación y el precio, los cafés con ratings más altos tienden a tener precios más elevados.

Además, la influencia del tipo de tostión es evidente, las tostiones "Medium-Light", "Medium" y "Light" son las que consistentemente alcanzan tanto los ratings más altos como los precios más elevados, sugiriendo que estos tipos de tostión están asociados con la excelencia y el valor en el mercado del café.

Por el contrario, las tostiones "Dark" y "Medium-Dark" muestran una clara tendencia a asociarse con ratings y precios más bajos. Esto respalda la hipótesis de que el nivel de tostión influye en el puntaje sensorial y el precio, y que existe una relación directa entre puntaje y precio.

Gráfico 7. Gráfico de paredes



**Resúmenes numéricos:**

En la parte de resúmenes numéricos se verán reflejados los análisis estadísticos, tales como la media, la cual es una herramienta fundamental para comprender el comportamiento general de los datos y facilitar la toma de decisiones basada en evidencia. Al representar un valor central, la media permite identificar patrones, comparar grupos y resumir grandes volúmenes de información de manera simple y efectiva.

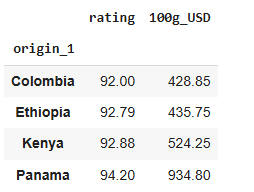
Media:

Este análisis cuantitativo de promedios valida la reputación de Colombia, Etiopía, Kenia y Panamá como productores de cafés con alto puntaje sensorial. Más importante aún, muestra una fuerte correlación positiva entre el rating promedio y el precio promedio dentro de este grupo de países destacados. Sin embargo, también subraya que la variable "País de Origen" no solo influye en la calidad, sino que también es un factor crítico que impulsa el precio, incluso entre cafés ya considerados de "alta calidad".

Muestra que el precio de Panamá es excepcionalmente alto, lo que sugiere la existencia de factores únicos (quizás el impacto del varietal Geisha) que elevan su valor de mercado mucho más allá de lo que implicaría solo el rating. Para un análisis más profundo, sería valioso examinar la distribución de los ratings y precios dentro de cada país y considerar otras variables como el varietal o el método de procesamiento.

Los cuatro países destacados (Colombia, Etiopía, Kenia, Panamá) muestran consistentemente ratings promedio muy altos, todos por encima de 92.00. Esto valida la premisa de la hipótesis de que estos países son reconocidos por producir cafés de alta calidad sensorial.

Tabla 1.



Correlación:

Este resultado de correlación cuantitativo es consistente con las observaciones visuales de los gráficos de dispersión que analizamos previamente. En esos gráficos, veíamos una tendencia ascendente (positiva) entre rating y precio, pero también una considerable dispersión de los puntos, especialmente en los ratings más altos. Esa dispersión es lo que se refleja en una correlación débil a moderada. Si la correlación hubiera sido muy fuerte (por ejemplo, 0.8 o 0.9), veríamos los puntos del gráfico de dispersión mucho más agrupados alrededor de una línea recta ascendente.

El coeficiente de correlación de 0.25 confirma una relación lineal positiva entre el rating y el precio, validando la dirección de la Hipótesis 3. Sin embargo, su baja magnitud indica que esta relación es débil.

**Conclusiones:**

Los resultados del análisis permiten confirmar las tres hipótesis planteadas, aunque con ciertas variaciones que evidencian la complejidad del mercado de cafés de especialidad y la influencia de factores adicionales.

Primero, se valida que los cafés con tostión media tienden a alcanzar puntajes sensoriales más altos en comparación con aquellos de tostión alta, sin embargo, se vieron mejores resultados con aquellos cafés de tostian Medium-light, esto sugiere que este nivel de tostión favorece el equilibrio de atributos como aroma, acidez y cuerpo, fundamentales para una buena calificación en cata.

En segundo lugar, se cumple la hipótesis de que los países de origen reconocidos por su alta calidad como Colombia, Etiopía, Panamá y Kenya registran en general mayores puntajes y precios, lo que refleja el prestigio y la reputación de estos orígenes en el mercado internacional. Esto confirma que el origen sigue siendo un factor determinante en la percepción y en si presencia de calidad, al igual que en el posicionamiento comercial del café.

También se respalda la hipótesis de que existe una relación entre el puntaje y el precio, ya que los cafés con ratings más bajos (<88) presentan precios consistentemente bajos. Sin embargo, se identifican excepciones notables en cafés con puntajes altos (>94) que no siempre alcanzan precios elevados, y en otros casos, cafés con puntajes medios (91–93) que presentan precios extraordinariamente altos. Estos casos evidencian que, aunque la calidad medida por el puntaje influye en el precio, no es el único determinante.

En síntesis, las hipótesis se confirman en términos generales, pero los resultados también revelan que existen otras variables no evaluadas como la exclusividad del lote, el tipo de proceso, el empaque, las certificaciones o las estrategias de marketing que también juegan un papel clave en la formación del precio final.

**Corrección:**

Buenas noches Valeria!!! ¡Muy original tu proyecto, como buen amante del café me encanto! Si bien tengo nociones generales en relación al proceso del café normalmente se desconocen muchos términos por lo que te invito a que detalles y coloques una breve descripción a las variables para que el que lo lea lo pueda entender e incluso te sirva a vos para un futuro. Me encanto la forma de presentación, la redacción y planteo de las hipótesis me van llevando por una historia muy bien contado (felicitaciones por el storytelling) Cundo hagas el análisis multivariado ten en cuenta que tienes que hablar de las tres variables relacionándose entre si, por ejemplo en el apartado de “Precio por tipo de tueste y país de origen” empiezas a relacionarlo con el raiting siendo que esta variable no aparece en el gráfico, y me da la impresión que dejas relegada la variable del país que creo que da muy buena información porque marca outliers muy claros y están dados porque en EEUU pareciera que están dispuestos a pagar mucho más por el café o bien en este país es más caro. ¡El siguiente gráfico de relación de precio y tipo de tueste si encaja perfectamente en el análisis realizado! ¡¡¡¡El trabajo esta super completo me gusto un montón el análisis, se nota el tiempo y esfuerzo dedicado!!!! ¡Sigue así que queda mucho por delante!