

Análisis del Dataset 'Flavors of cacao'

Naturaleza del dato y técnicas de análisis aplicadas

Revisión documental

- Origen: Kaggle (Rachael Tatman), datos recopilados por Brady Brelinski.
- Dataset: 1795 muestras de barras de chocolate oscuro.
- Es un dataset de evaluación experta con información sensorial y geográfica.

Clasificación de variables

Variable	Tipo	Descripción	¿Qué aporta?
company	Cualitativa, Nominal	Nombre de la empresa que produce la barra	Puede asociarse con la calidad promedio o reputación de la empresa.
specific_bean_origin_or_bar_name	Cualitativa, Nominal	Región o variedad específica del grano	Podría influir en el sabor o calidad del producto.
ref	Cuantitativa, Discreta	Número de referencia interno	No aporta información relevante por sí solo.
review_date	Cuantitativa, Discreta	Año en el que se realizó la evaluación	Permite analizar cambios de tendencia a lo largo del tiempo.
cocoa_percent	Cuantitativa, Continua	Porcentaje de cacao en la barra	Se asocia con sabor, amargor o calidad.
company_location	Cualitativa, Nominal	País donde está ubicada la empresa productora	Puede influir en el estilo de producción o calidad regional.
rating	Cuantitativa, Continua	Valoración otorgada por expertos (de 1 a 5)	Variable objetivo: refleja la calidad percibida del chocolate.
Bean type	Cualitativa, Nominal	Tipo de especie del grano utilizado	Afecta sabor, textura o calidad final.
Broad bean origin	Cualitativa, Nominal	País de origen del grano de cacao	El origen agrícola puede afectar calidad y perfil del sabor.

Limpieza

- Modificamos la variable de cocoa_percent para que fuera un float.
- Corregimos/limpiamos la var 'Bean Type' para reunificar valores y eliminar especificaciones de la subespecie del grano.
 - Ej: Trinitario (arriba) => Trinitario;
- La var 'Broad Bean Origin' tenía muchos errores gramaticales, inserción rápida sin comprobación...
 - Ej; Dom. Rep. || R.D. || dominican rep. ==> Dominican Republic ; guat. => Guatemala
- Las vars 'Broad Bean Origin' y 'Company location' las subdividimos en continentes, generando dos columnas extra con más datos.

Data columns (total 9 columns):			
#	Column	Non-Null Count	Dtype
0	company	1795 non-null	object
1	specific_bean_origin	1795 non-null	object
2	ref	1795 non-null	int64
3	review_date	1795 non-null	int64
4	cocoa_percent	1795 non-null	object
5	company_location	1795 non-null	object
6	rating	1795 non-null	float64
7	bean_type	1794 non-null	object
8	broad_bean_origin	1794 non-null	object

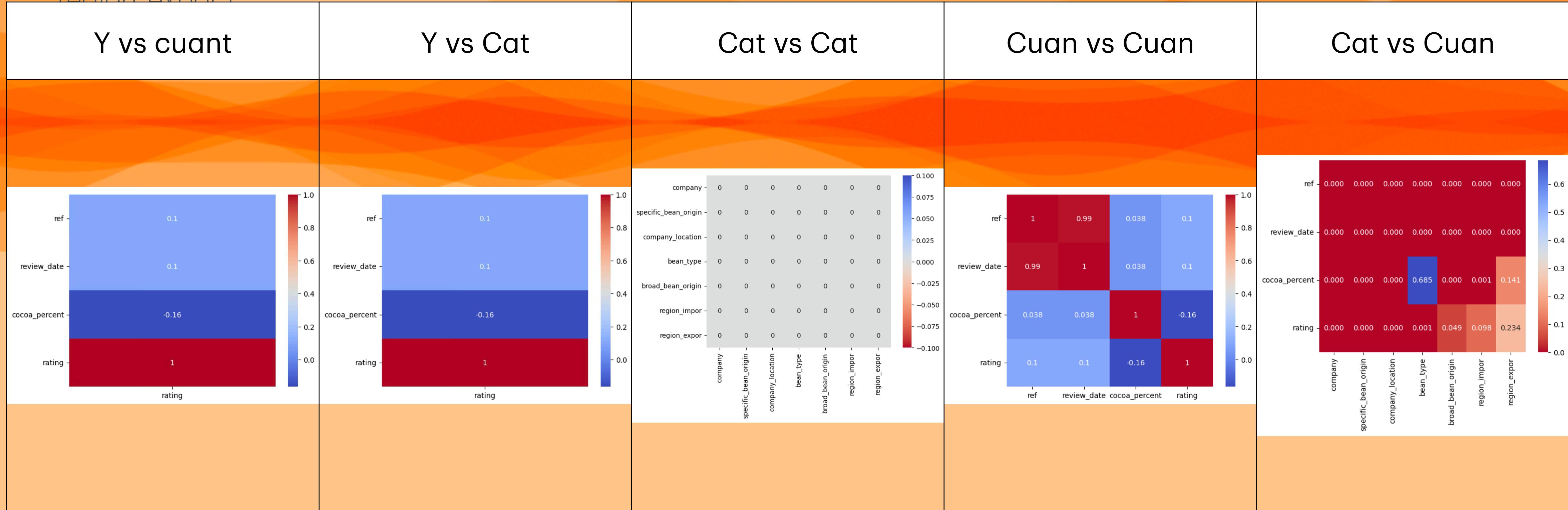
dtypes: float64(1), int64(2), object(6)

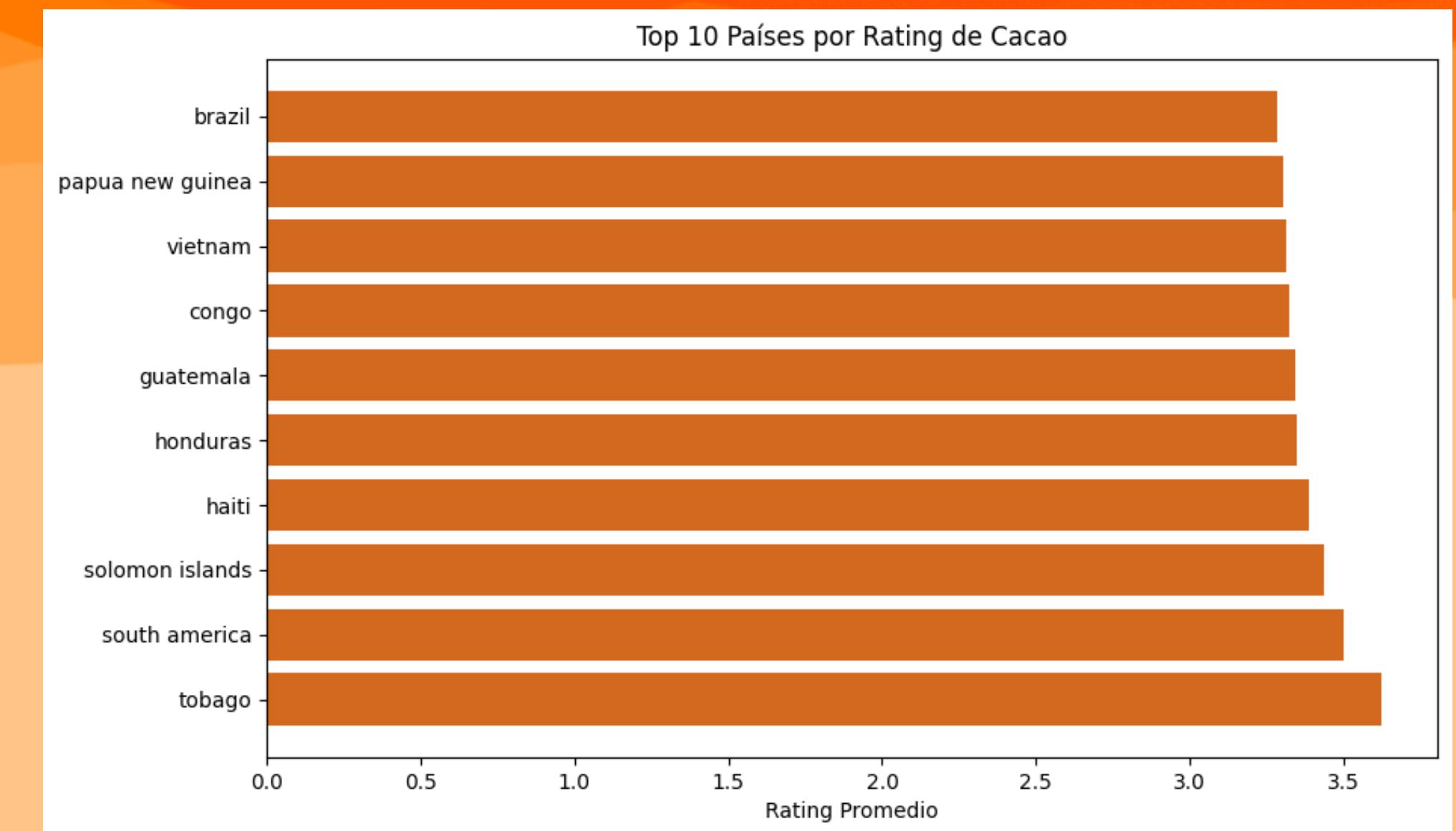
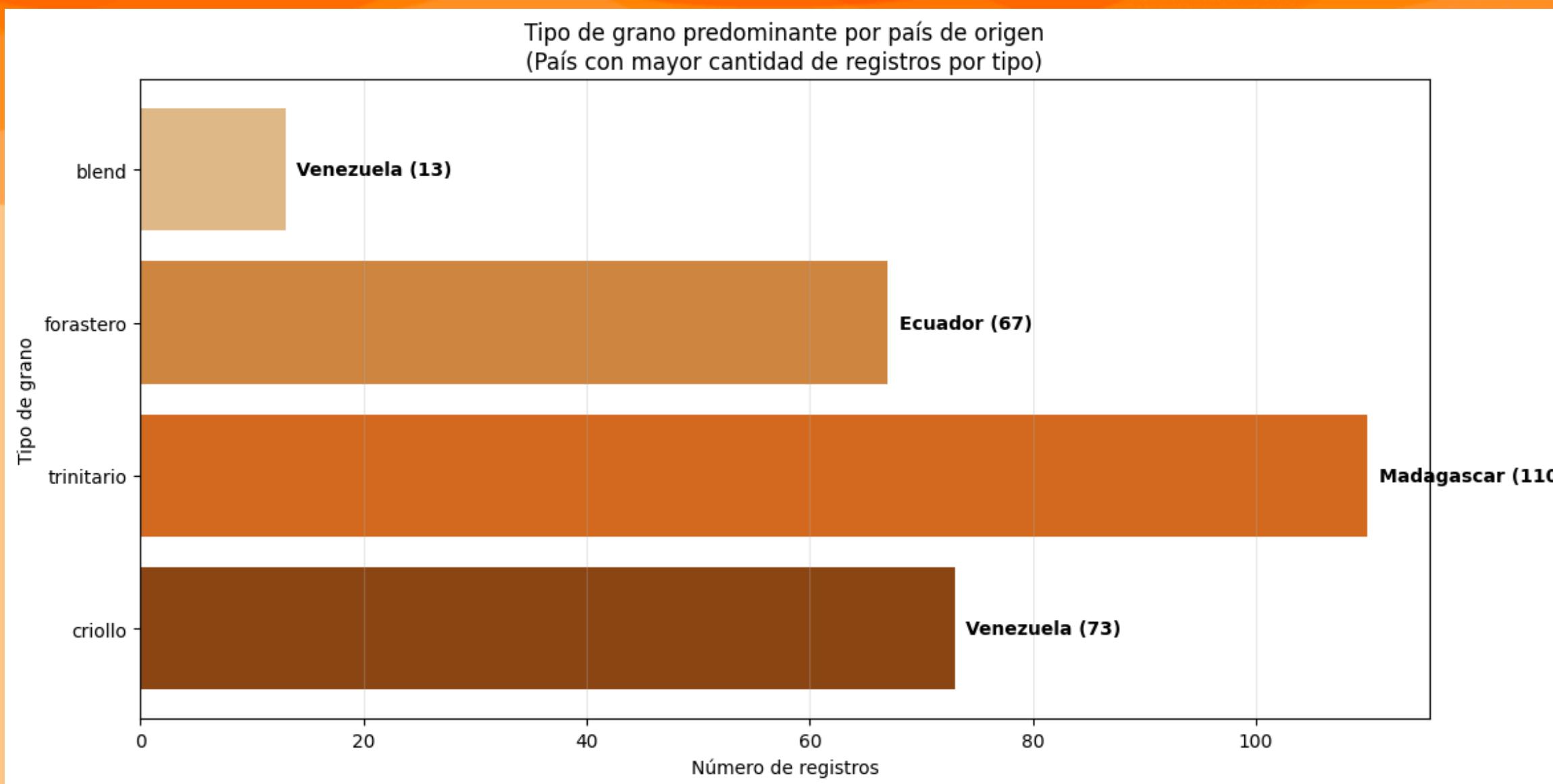
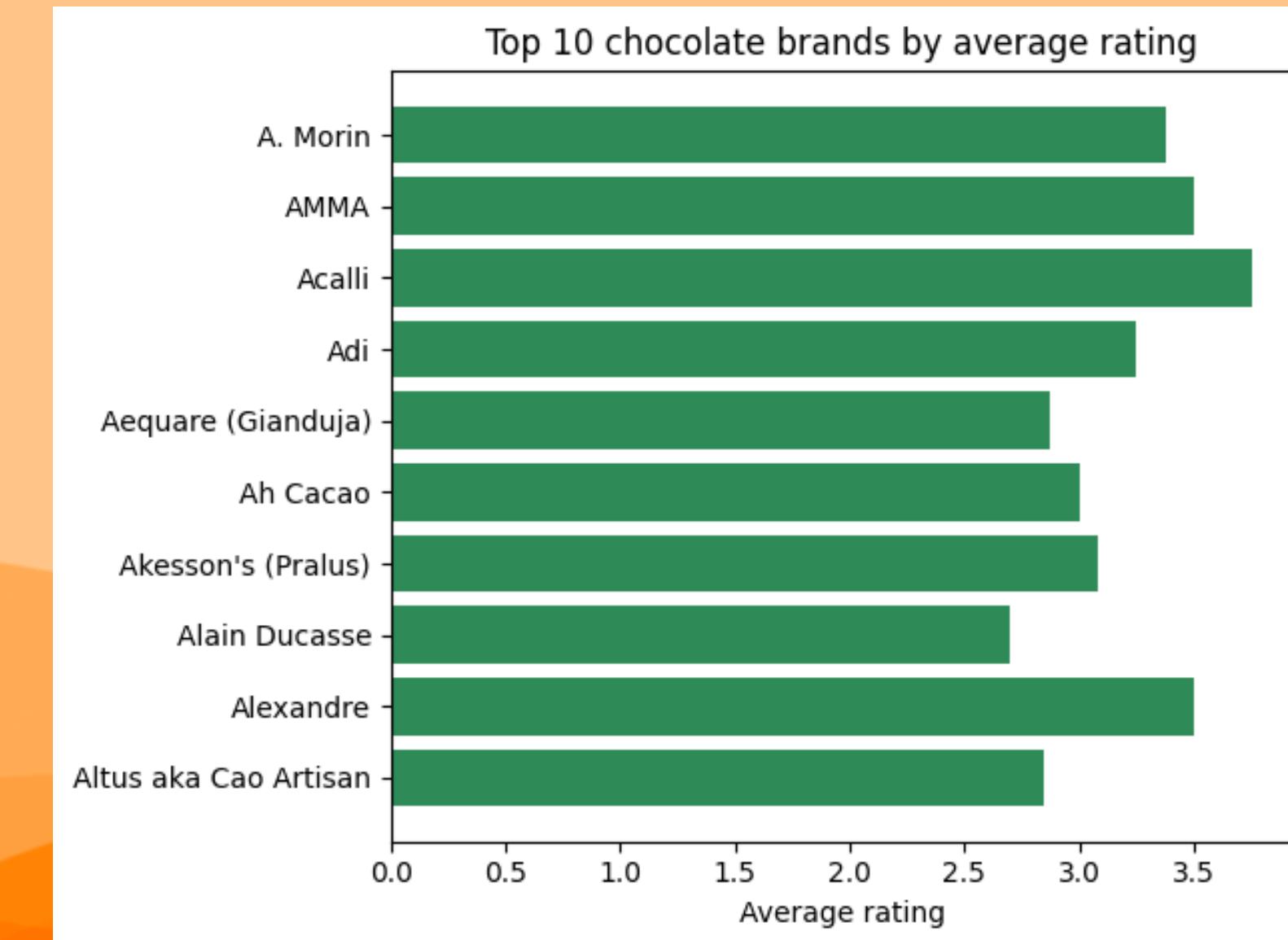
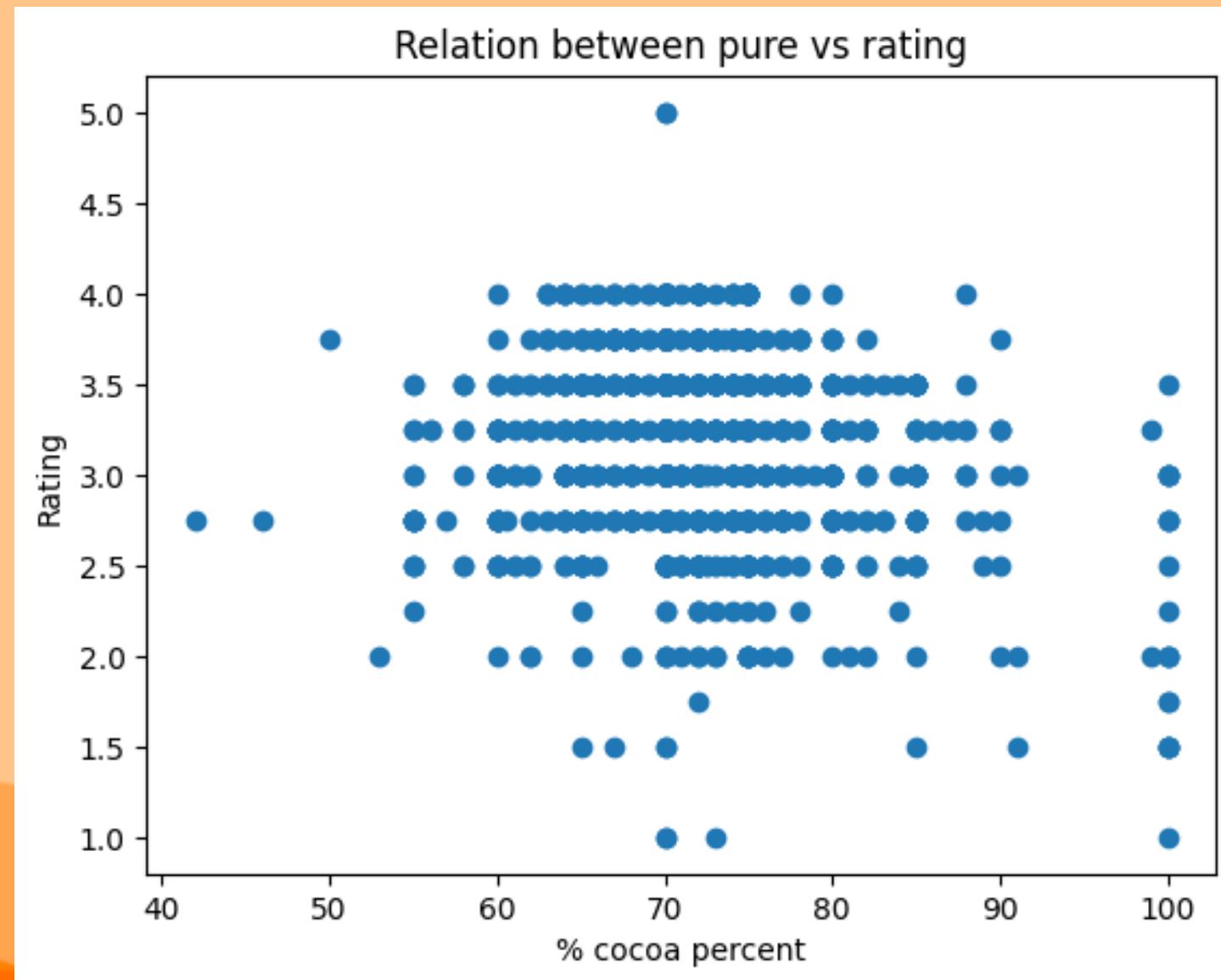
```
Data columns (total 11 columns):  
#   Column           Non-Null Count  Dtype     
---    
continent_dict = {  
    'north_america': [  
        'u.s.a.', 'canada', 'mexico', 'puerto rico', 'hawaii'  
    ],  
    'central_america': [  
        'guatemala', 'costa rica', 'nicaragua', 'honduras',  
        'belize', 'panama', 'el salvador'  
    ],  
    'south_america': [  
        'ecuador', 'colombia', 'venezuela', 'brazil', 'peru',  
        'argentina', 'bolivia', 'chile', 'suriname', 'south america'  
    ],  
    'europe_west': [  
        'france', 'u.k.', 'italy', 'belgium', 'switzerland', 'germany',  
        'austria', 'spain', 'denmark', 'netherlands', 'ireland',  
        'portugal', 'iceland', 'sweden', 'finland'  
    ]  
}
```

Análisis exploratorio

Selección propia de variables

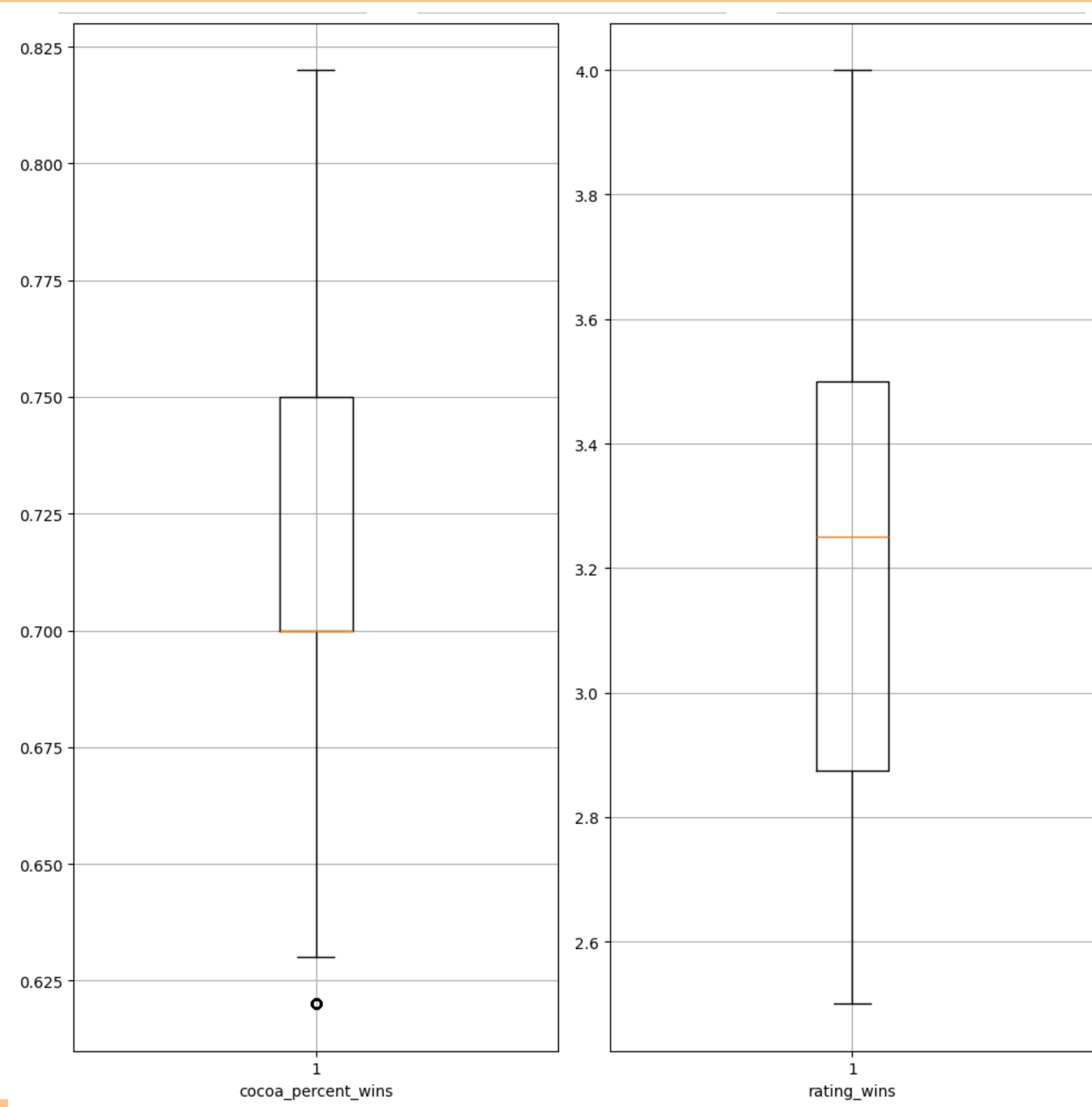
- Categóricas = ['company', 'specific(bean)_origin', 'company_location', 'bean_type', 'broad(bean)_origin', 'region_impor', 'region_expor']





Procesamiento

- Aplicamos una transformación que hacían ponderar



ado que vimos que

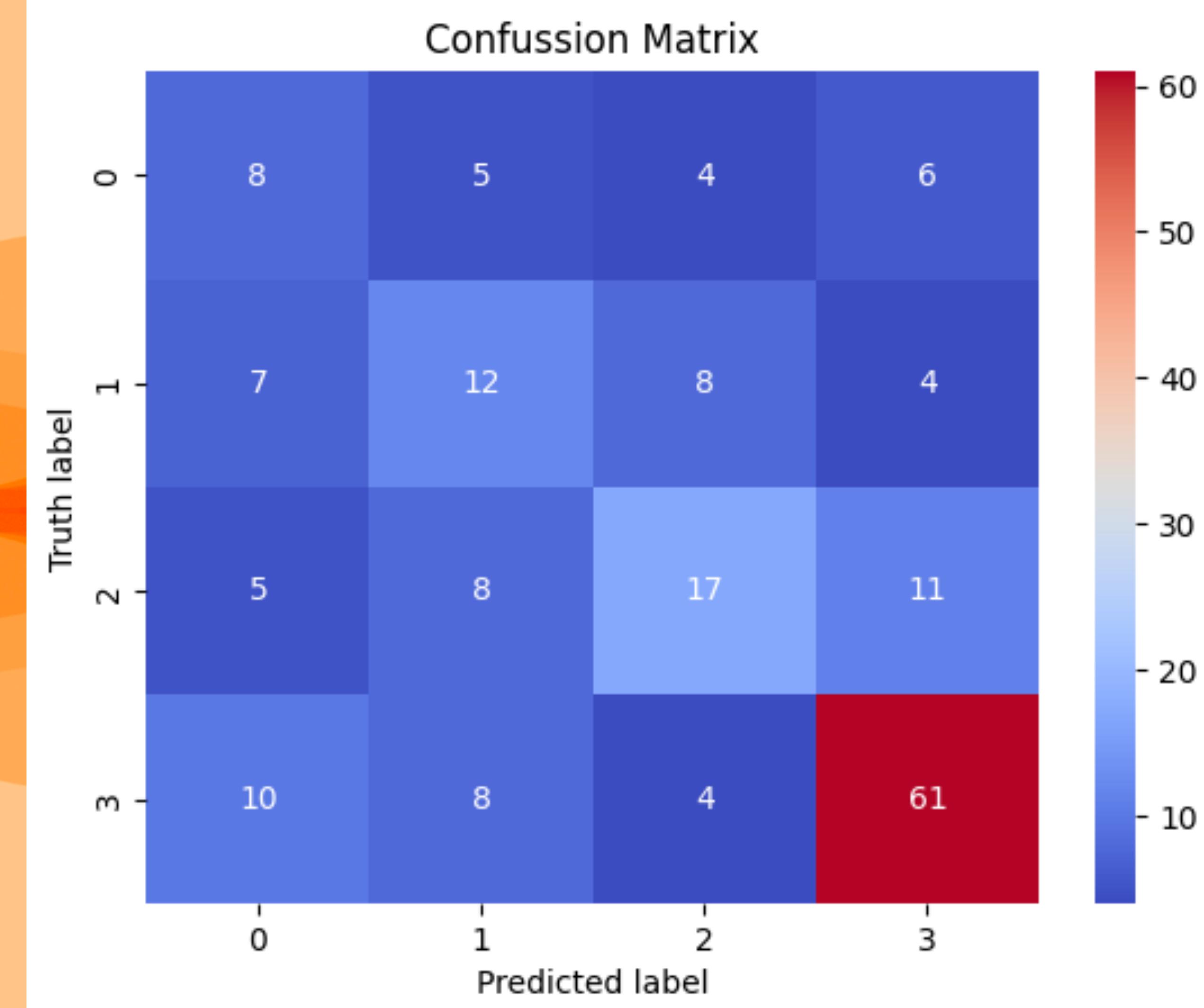
Procesamiento missings

- Teníamos valores faltantes tanto en bean_type (908/1795) y en broad_bean_origin (95/1795).
- Para poder eliminarlos, tenían que representar < 54 rows/columna (3% 1795) y < 718 del total (5% de 14360).

Random forest Classifier

'bean_type'

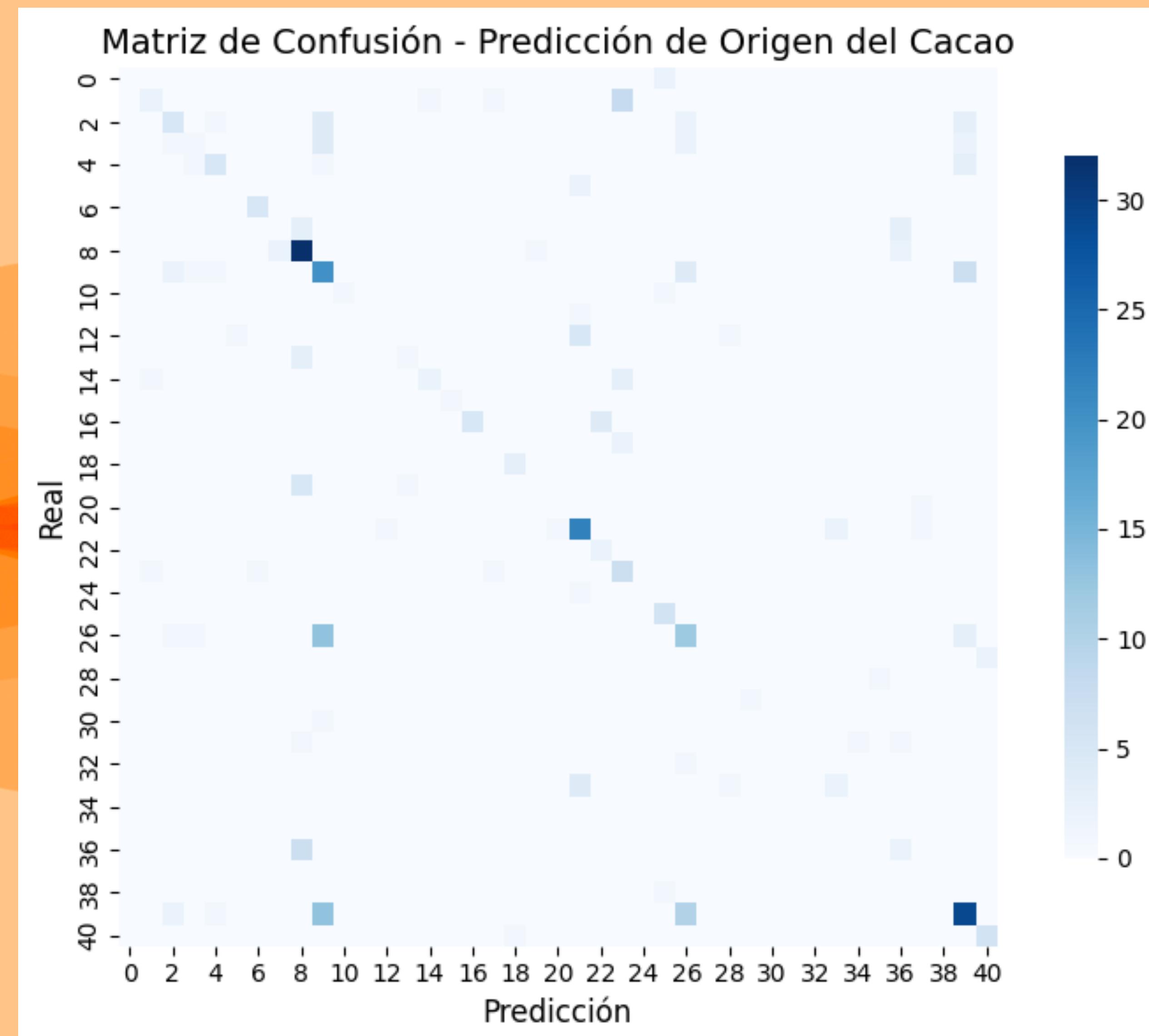
	precision	recall	f1-score	support
blend	0.27	0.35	0.30	23
criollo	0.36	0.39	0.38	31
forastero	0.52	0.41	0.46	41
trinitario	0.74	0.73	0.74	83
accuracy			0.55	178
macro avg	0.47	0.47	0.47	178
weighted avg	0.56	0.55	0.55	178



Random forest Classifier

'broad_beans_origin'

```
==== RESULTADOS DEL MODELO ====  
Accuracy: 0.491  
Precision: 0.475  
Recall: 0.491  
F1-Score: 0.465  
Países únicos en test: 39  
Países predichos: 35
```



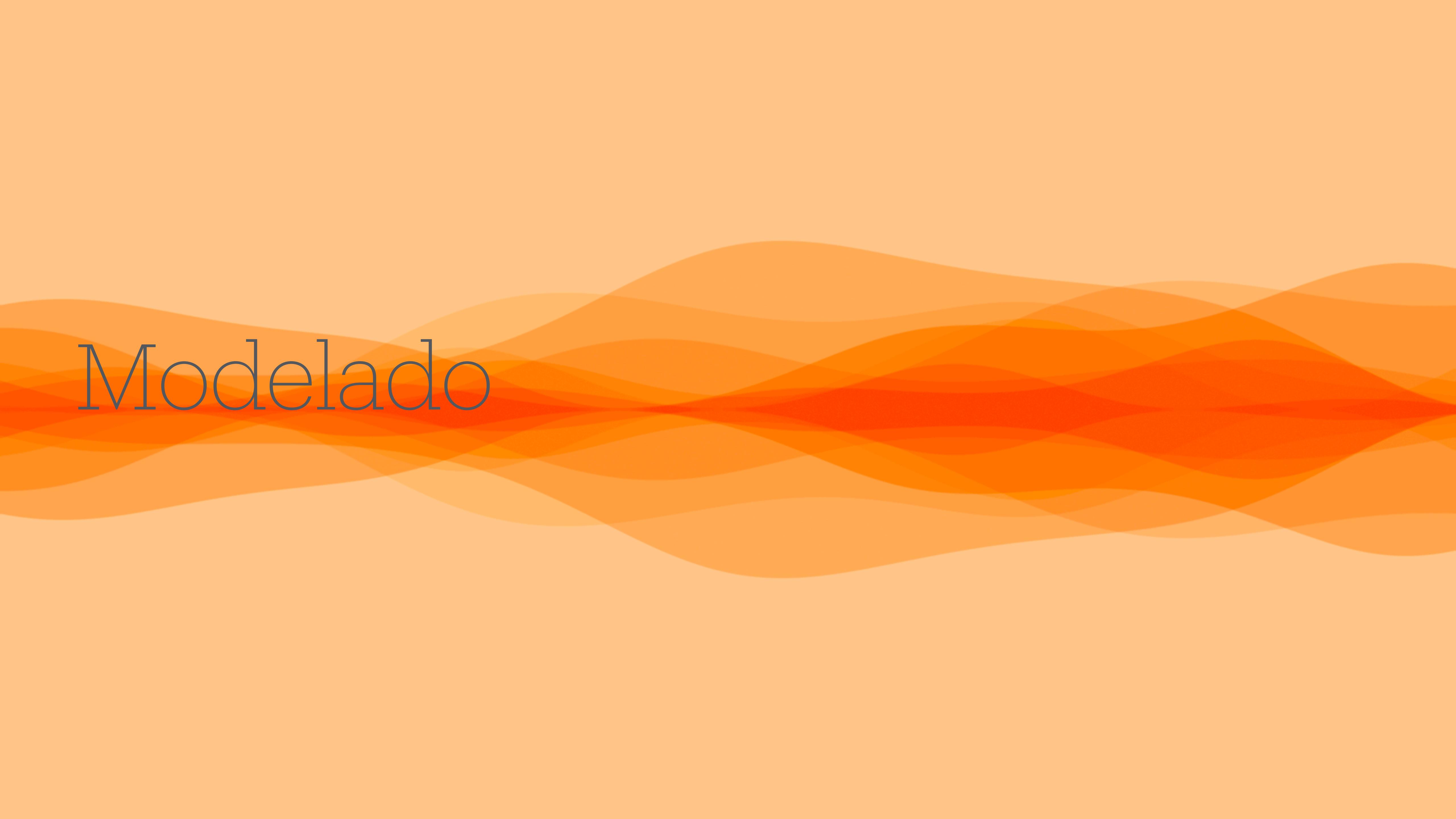
Selección de variables

RFE

- Se utilizó para reducir la dimensionalidad y centrarnos en las variables más relevantes que nos da este método.
- Selecciona de forma automática las características más importantes y va eliminando las que aportan menos valor predictivo.
- Company_location, bean_type y region_impor

OLS

- Primero con las variables resultantes del RFE, aplicamos el OLS, este modelo permite analizar cómo cada variable incide en la calificación, midiendo su significancia estadística (p-valores) y el impacto en la variable dependiente.
- Esta combinación permitió construir un modelo simple pero explicativo, centrado en las variables más influyentes según el análisis previo.
- Cocoa_percent, company_location, bean_type y broad_bean_origin.

The background features a series of overlapping, wavy layers in shades of orange, yellow, and cream, creating a sense of depth and motion.

Modelado

Intento de reducción con PCA

Sobre la variable 'company'

- La variable contenía 416 valores únicos.

¿ Por qué lo intentamos?

- Reducir el número de categorías.
- Capturar patrones o agrupamientos de compañías.
- Disminuir el ruido y la carga en los modelos.

¿Qué ocurrió?

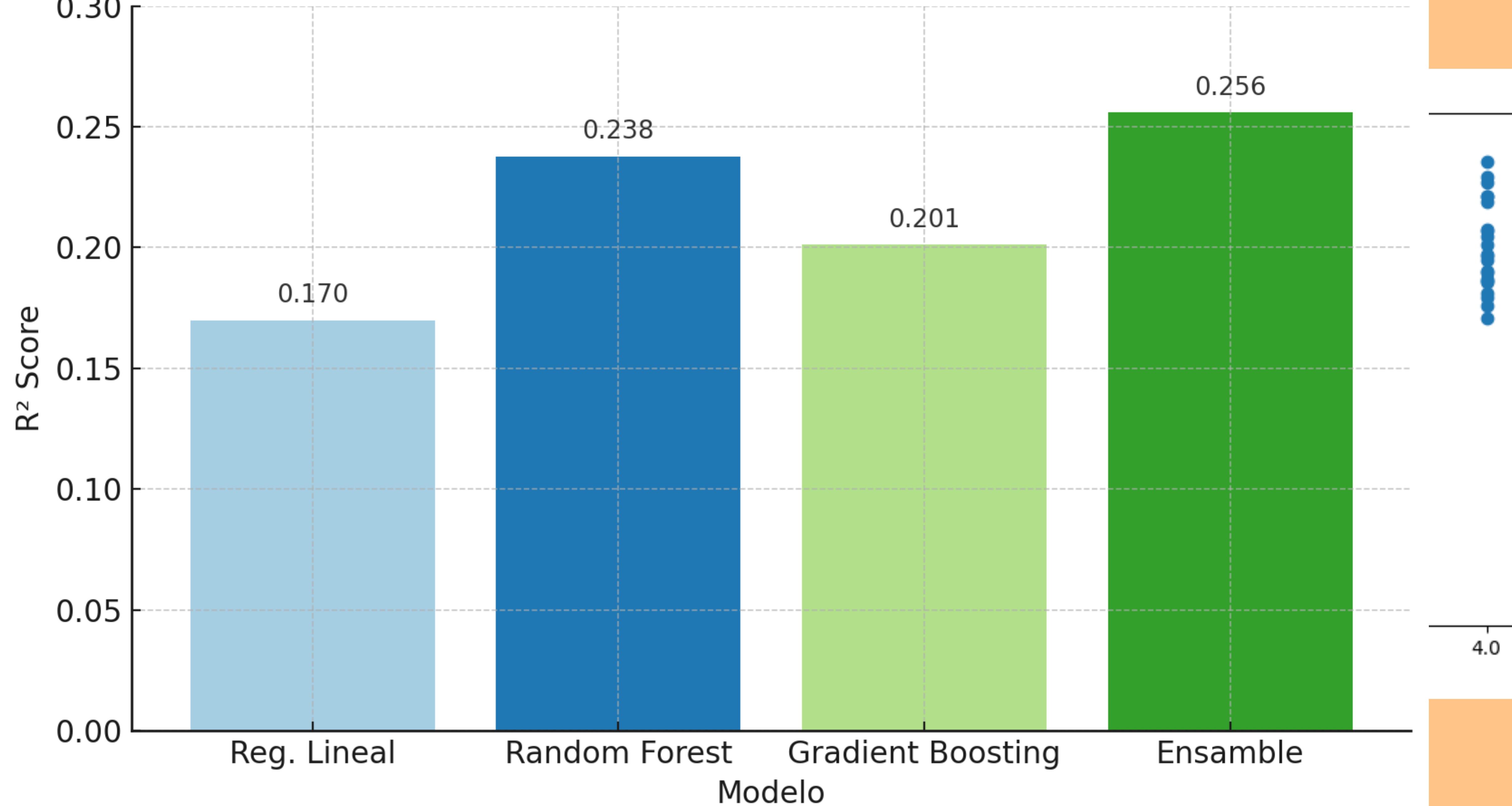
- Es una categoría nominal sin relación ordinal o numérica.
- Al aplicar el OneHotEncoding se generaron cientos de columnas y PCA no fue capaz de encontrar relación que explicaran la varianza.
- El resultado fue un modelo poco interpretable y sin mejora en el rendimiento predictivo.

Ensemble de modelo predictivo

LR, RFR y GBR

- Se entrenaron 3 modelos distintos optimizados individualmente con GridSearchCV.
- Luego, se combinaron para crear un modelo de ensamble con Voting Regressor, con el objetivo de:
 - Aprovechar lo mejor de cada modelo.
 - Mejorar la robustez y estabilidad de las predicciones.
 - Reducir errores individuales de cada estimador.
- Variables utilizadas:
 - Categóricas: bean_type, region_impor, region_expor y company
 - Numéricas: review_date y cocoa_percent_wins
- El modelo ensemble mostró mejor rendimiento global que cualquier otro modelo individual.

Comparación de R² entre Modelos (Conjunto de Prueba)



Resultados clave y conclusiones

- **Países destacados (company_location)**

- Países más frecuentes: Estados Unidos, Francia y Canadá concentran la mayor cantidad de fabricantes.
- Mejores valoraciones promedio: Suiza, Bélgica y Reino Unido sobresalen por calidad percibida en las evaluaciones.
- Conclusión: La localización de la empresa puede influir en la percepción de calidad del chocolate.

- **Relación entre % cacao y rating:**

- Se observó una tendencia negativa leve; a medida de que aumenta el % del cacao, la calificación tiende a disminuir ligeramente.
- Posible causa: mayor amargor percibido o preferencias de sabores más suaves.
- Conclusión: el % del cacao no garantiza una mejor calificación.

- **Tipo de grano más influyente:**

- Tipos más comunes: Trinitario, Forastero y Criollo.
- Mejor puntuado: en general, el grano Criollo obtuvo mejores valoraciones medias, aunque es menos frecuente.
- Conclusión: La especie del grano es una variable relevante y marca la diferencia en la calidad.

The background features a series of overlapping, wavy layers in shades of orange, yellow, and cream, creating a sense of depth and motion.

Preguntas

Análisis del Dataset ‘Flavors of cacao’

Naturaleza del dato y técnicas de análisis aplicadas

Gracias por vuestra atención

Asier Novio y Martí Vallhonrat - Julio 2025