

## Resumen del Notebook — Ramen Ratings

### INTRODUCCIÓN

El conjunto de datos Ramen Ratings reúne más de 2.500 reseñas de fideos instantáneos registradas en “The Ramen Rater”. Cada fila es una evaluación con variables como marca, variedad (Variety), país, estilo (Style) y calificación (Stars).

### HIPÓTESIS

**H1 — País y calidad:** Algunos países (ej. Japón, Corea, Singapur) presentan medias de Stars diferentes.

**H2 — Estilo (Cup / Bowl / Pack):** El Style (Cup/Bowl/Pack/Tray) no afecta la puntuación.

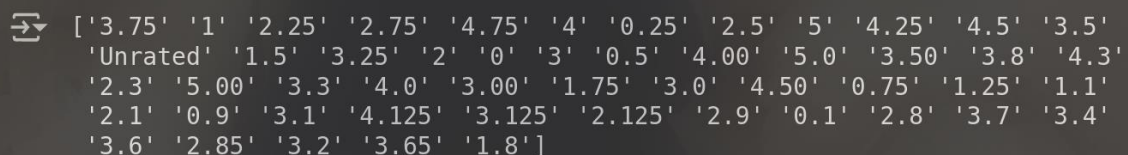
### WORKFLOW

Flujo propuesto: 1) limpieza y normalización; 2) codificación de categóricas; 3) creación de estadísticas por marca/país; 4) modelado (regresión/ordinal); 5) evaluación por validación cruzada.

### MANEJO DE NULOS Y PRE-PROCESADO

Se detectaron valores especiales en 'Stars' (ej. 'Unrated') y algunos NaN en 'Top Ten' y 'Style'. Decisión tomada en el notebook: convertir 'Stars' a numérico y eliminar filas sin calificación (pocas), y eliminar las pocas filas con 'Style' NaN.

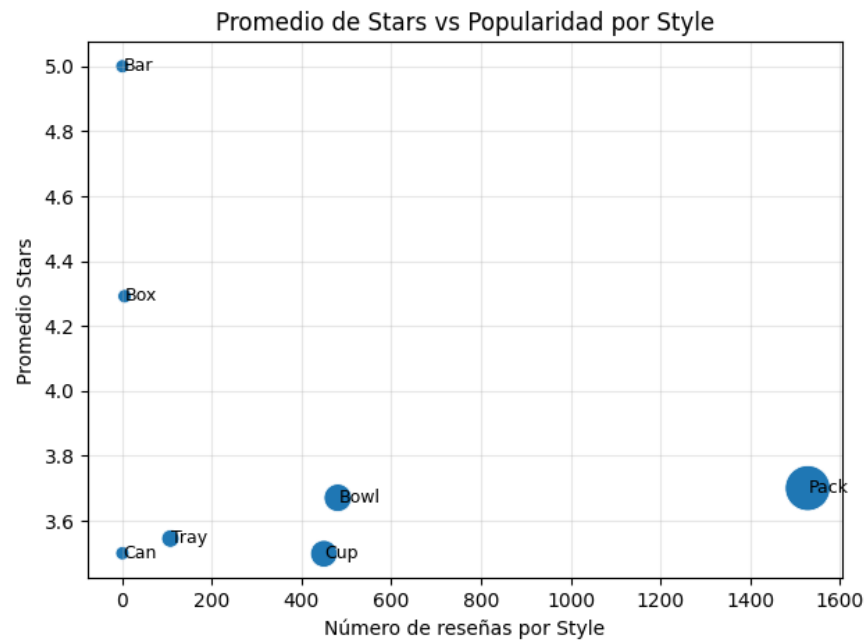
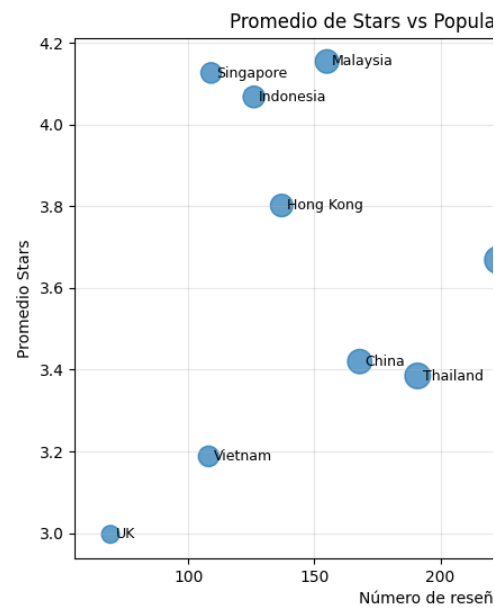
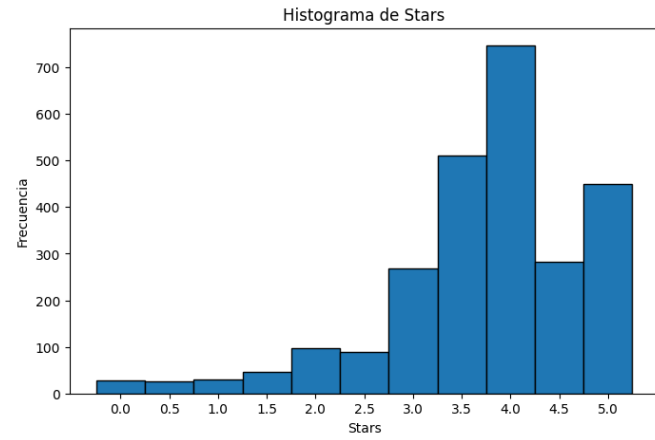
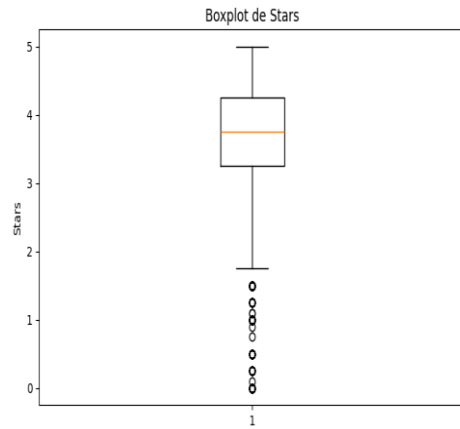
```
[ ] print(df['Stars'].unique())
```

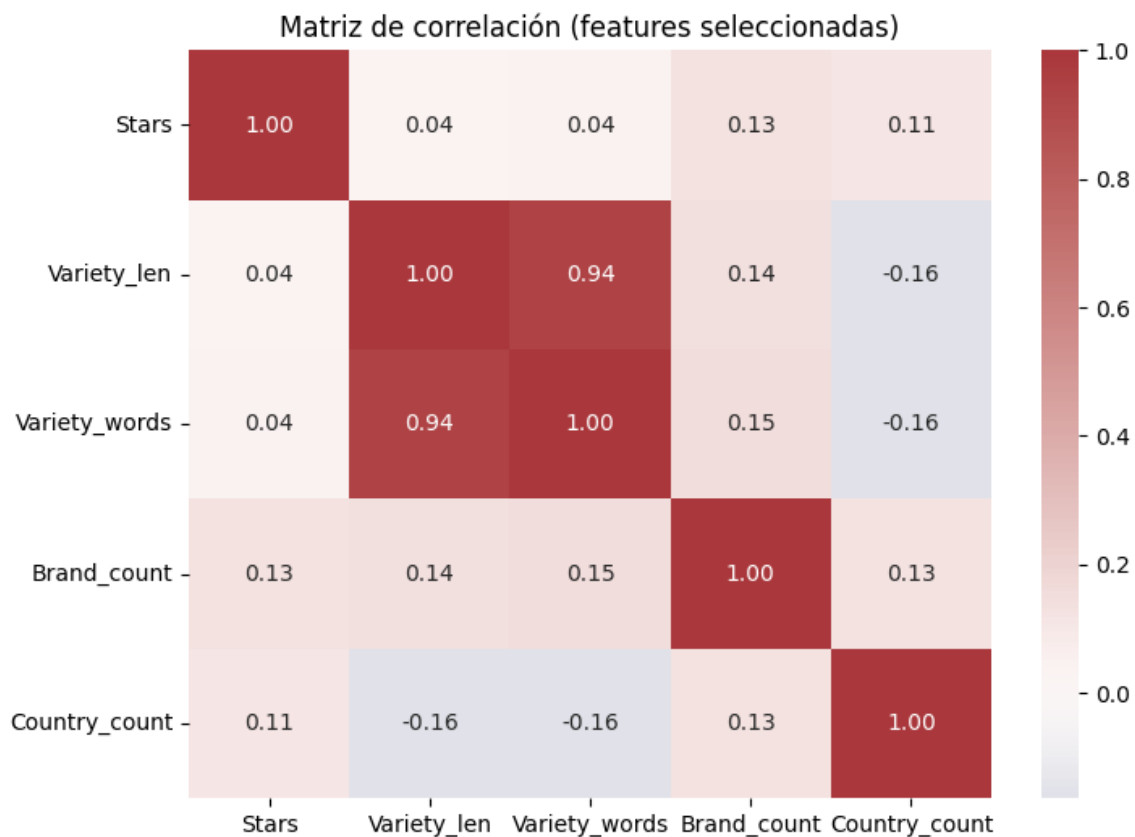


```
['3.75' '1' '2.25' '2.75' '4.75' '4' '0.25' '2.5' '5' '4.25' '4.5' '3.5'
'Unrated' '1.5' '3.25' '2' '0' '3' '0.5' '4.00' '5.0' '3.50' '3.8' '4.3'
'2.3' '5.00' '3.3' '4.0' '3.00' '1.75' '3.0' '4.50' '0.75' '1.25' '1.1'
'2.1' '0.9' '3.1' '4.125' '3.125' '2.125' '2.9' '0.1' '2.8' '3.7' '3.4'
'3.6' '2.85' '3.2' '3.65' '1.8']
```

### VISUALIZACIONES

Se generaron visualizaciones para explorar la distribución y relaciones: histograma, boxplot, scatter por país (Count vs MeanStars), scatter por Style, y heatmap de correlación con features derivadas.





### OBSERVACIONES (EDA)

- La distribución de 'Stars' está sesgada hacia la derecha (más valores altos).
- No hay valores fuera de rango (0-5). Los valores extremos (1 y 2) son raros.
- Variación por país: algunos países muestran promedios cercanos a 5.
- Pocas correlaciones entre las features numéricas simples y 'Stars' (posible limitación de features).

### MODELADO

Se utilizó una regresión Ridge (pipeline con StandardScaler para numéricas y OneHotEncoder para categóricas). Features: 'Variety\_len', 'Variety\_words', 'Brand\_count', 'Country\_count', más 'Brand', 'Country', 'Style' como categóricas.

```
# --- Pipeline / Preprocessor ---
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), num_features),
        ('cat', ohe, cat_features)
    ],
    remainder='drop'
)

pipe = Pipeline([
    ('pre', preprocessor),
    ('model', Ridge(alpha=1.0, random_state=42))
])
```

## EVALUACIÓN (TEST)

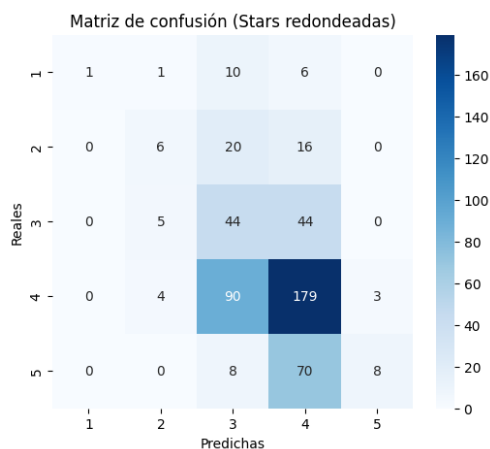
Métricas calculadas en el set de test:

RMSE: 0.9235

MAE : 0.6776

R2 : 0.1503

Además se discretizaron predicciones redondeando a 1..5 para generar una matriz de confusión y un reporte de clasificación.



## CONCLUSIONES

- El modelo entrega predicciones aproximadas (MAE < 1 en el notebook), pero  $R^2$  bajo y accuracy < 50% indican poca capacidad para explicar la varianza y predecir exactamente.
- Sesgo hacia la clase mayoritaria (4 estrellas): el modelo tiende a predecir la clase más

frecuente.

- Falta de features relevantes: se sugiere enriquecer con atributos del texto (NLP sobre 'Variety'), ingredientes, información nutricional si estuviera, o datos del autor/fecha de la reseña.