Universidad del Desarrollo
Facultad de Ingeniería
dataScience

Tarea 2 - Análisis de Datos
Regresión Lineal Simple

"Most Streamed Spotify Songs 2024"

Profesor: Boris Panes

Integrantes:
Bruno San Martín
Carlos Saquel
Camilo Rivera
Juan Ugalde

Domingo 6 de octubre de 2024

Introducción

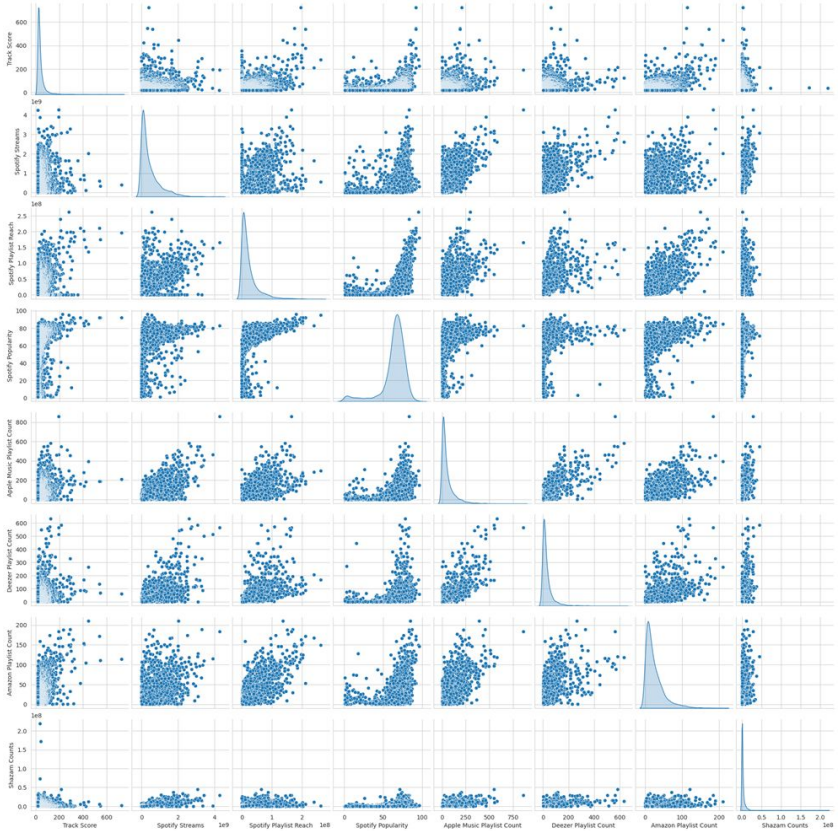
En esta segunda entrega, se llevará a cabo una regresión lineal utilizando el mismo dataset "Most Streamed Spotify Songs 2024". El objetivo es analizar la relación entre diferentes variables del conjunto de datos, en búsqueda de factores como las reproducciones y los likes influyen en el éxito y popularidad de las canciones en diversas plataformas.

Link del dataset:
https://www.kaggle.com/datasets/nelgiriyewithana/most-streamed-spotify-songs-2024

Selección de variables
Análisis del contenido y distribución

Para seleccionar las variables independientes (X) y dependientes (Y) de la regresión lineal, se realizó un análisis del contenido y la distribución de los datos. A través de gráficos de dispersión, se evaluaron visualmente las relaciones entre las variables, identificando aquellas que mostraban una posible correlación.

En el contexto de este ejercicio, resulta interesante analizar cómo la plataforma **Shazam, diseñada para identificar canciones específicas**, se relaciona con la popularidad de esas **canciones en Spotify**. Esto sugiere que los usuarios que utilizan Shazam para descubrir una canción podrían, posteriormente, reproducirla en Spotify, estableciendo un vínculo directo entre el reconocimiento musical y su consumo en la plataforma de streaming.



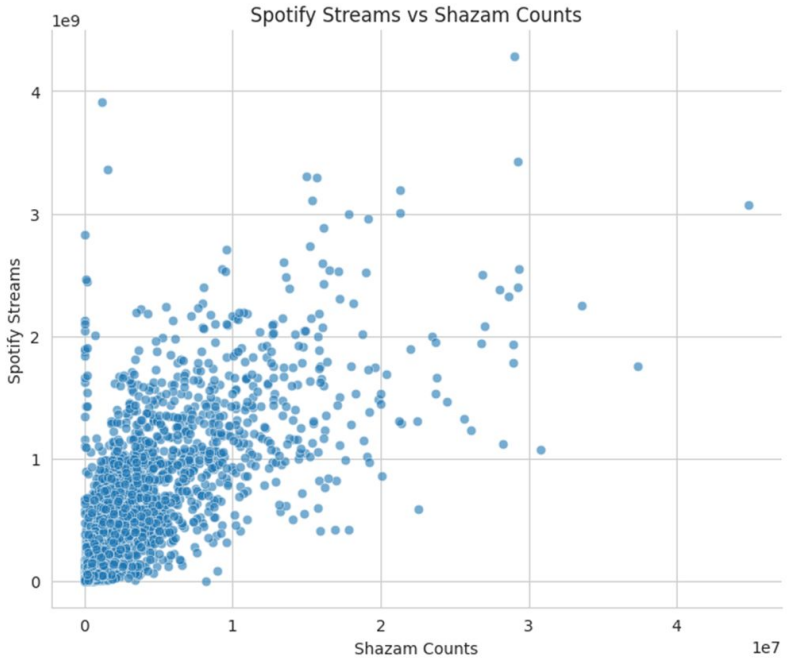
Selección de variables
Reproducciones en Spotify vs Recuento en Shazam

Este enfoque permitió seleccionar las búsquedas en Shazam como variable independiente (X) y las reproducciones en Spotify como variable dependiente (Y), dado que el gráfico de dispersión sugiere una relación significativa entre ambas variables.

```
# Correlación entre 'Shazam Counts' y 'Spotify Streams'
print(spotify_data_clean['Shazam Counts'].corr(spotify_data_clean['Spotify Streams']))
```

```
0.7346559010373636
```

La correlación entre las dos variables es de 0.73, lo que indica una fuerte relación positiva entre las búsquedas en Shazam y la popularidad de las canciones en Spotify, sugiriendo que a medida que una canción es identificada más veces en Shazam, también tiende a ser reproducida más en Spotify.



Cálculo de los coeficientes

```
# Definir X (Shazam Counts) y Y (Spotify Streams)
X = spotify_data_clean['Shazam Counts']
Y = spotify_data_clean['Spotify Streams']
```

Cálculo del R^2 $R^2 = \frac{SE}{ST}$

```
# Calcular la media de Y
Y_mean = np.mean(Y)

# Calcular SST (Suma total al cuadrado)
SST = np.sum((Y - Y_mean) ** 2)

# Calcular RSS (Suma de residuos al cuadrado)
RSS = np.sum((Y - predictions) ** 2)

# Paso 4: Calcular R^2
R_squared = 1 - (RSS / SST)

print("SST:", SST)
print("RSS:", RSS)
print("R^2:", R_squared)
```

```
SST: 1.0354723820969028e+21
RSS: 4.7660796018403384e+20
R^2: 0.5397192929290205
```

Cálculo de los coeficientes de la regresión $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

```
# Cálculo manual de los coeficientes
# Número de observaciones
n = len(X)

# Sumas necesarias
sum_X = X.sum()
sum_Y = Y.sum()
sum_XY = (X * Y).sum()
sum_X_squared = (X ** 2).sum()

# Cálculo de beta1 (pendiente)
beta1 = (n * sum_XY - sum_X * sum_Y) / (n * sum_X_squared - sum_X**2)

# Cálculo de beta0 (intercepto)
mean_X = sum_X / n
mean_Y = sum_Y / n
beta0 = mean_Y - beta1 * mean_X

print("Coeficiente beta1 (pendiente):", beta1)
print("Coeficiente beta0 (intercepto):", beta0)
```

```
Coeficiente beta1 (pendiente): 93.76184210509393
Coeficiente beta0 (intercepto): 207633648.27422264
```

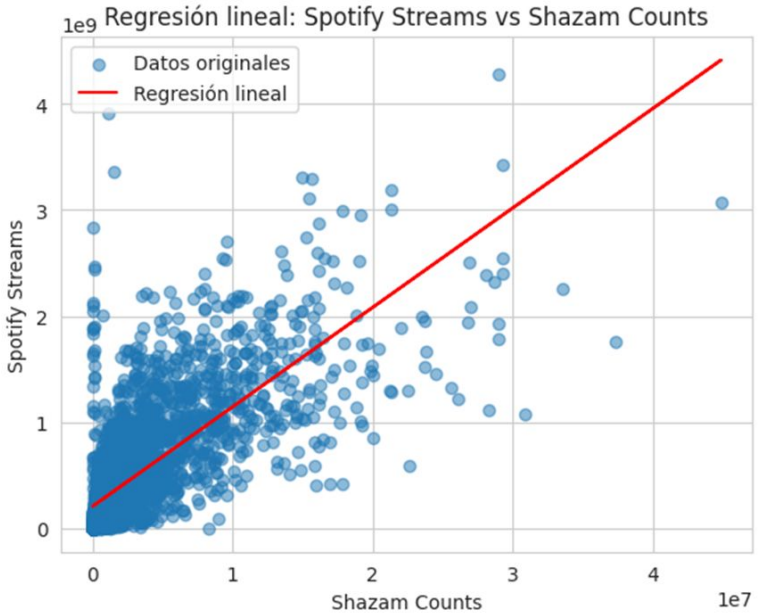
Regresión Lineal Simple

OLS Regression Results

Dep. Variable:	Spotify Streams	R-squared:	0.540
Model:	OLS	Adj. R-squared:	0.540
Method:	Least Squares	F-statistic:	4655.
Date:	Fri, 04 Oct 2024	Prob (F-statistic):	0.00
Time:	12:58:21	Log-Likelihood:	-83738.
No. Observations:	3972	AIC:	1.675e+05
Df Residuals:	3970	BIC:	1.675e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.076e+08	6.42e+06	32.342	0.000	1.95e+08	2.2e+08
Shazam Counts	93.7618	1.374	68.229	0.000	91.068	96.456

Omnibus:	1701.859	Durbin-Watson:	1.819
Prob(Omnibus):	0.000	Jarque-Bera (JB):	19887.729
Skew:	1.717	Prob(JB):	0.00
Kurtosis:	13.410	Cond. No.	5.46e+06



Resultados

R-squared (R-cuadrado): 0.540 Interpretación: El valor de R-cuadrado indica qué porcentaje de la variación en la variable dependiente (Spotify Streams) puede explicarse por la variable independiente (Shazam Counts). En este caso, un valor de 0.540 significa que el 54% de la variación en las reproducciones de Spotify puede explicarse por las búsquedas en Shazam. Es un valor moderado, lo que sugiere que hay una relación importante entre ambas variables, aunque no es perfecta.

$$R^2 = \frac{SE}{ST}$$

Coeficiente de Shazam Counts: 93.7618 Interpretación: Este coeficiente indica que, en promedio, por cada búsqueda adicional en Shazam, se asocian aproximadamente 93.76 reproducciones adicionales en Spotify. Es decir, la relación es positiva, y un aumento en las búsquedas en Shazam lleva a un aumento en las reproducciones en Spotify.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Resultados

Intercepto (const): 2.076e+08 Interpretación: El valor del intercepto (207,600,000) representa el valor estimado de las reproducciones de Spotify cuando las búsquedas en Shazam son cero. Este valor es bastante alto, lo que indica que incluso sin búsquedas en Shazam, se esperarían en promedio 207 millones de reproducciones de Spotify para las canciones de este conjunto de datos.

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}$$

P-valor de Shazam Counts: 0.000 Interpretación: El p-valor es prácticamente cero, lo que significa que hay evidencia suficiente para rechazar la hipótesis nula, que establece el coeficiente es igual a 0. Esto implica que existe una relación estadísticamente significativa entre las búsquedas en Shazam y las reproducciones en Spotify. En otras palabras, es muy poco probable que esta relación ocurra por azar.

F-statistic: 4655 Interpretación: El estadístico F es utilizado para evaluar la hipótesis de que el modelo tiene valor predictivo. Un valor alto como este, junto con un p-valor asociado de 0.000, indica que el modelo en su conjunto es altamente significativo.

Alternativa con Transformación Logarítmica

Considerando la escala de los datos es una buena alternativa probar con una transformación logarítmica:

$$Y = \log(Y_{ini} + 1) \quad X = \log(X_{ini} + 1)$$
$$Y = \beta_0 + \beta_1 X$$

```
# Correlación entre 'Shazam Counts Log' y 'Spotify Streams Log'
print(spotify_data_clean_log['Shazam Counts Log'].corr(spotify_data_clean_log['Spotify Streams Log']))
```

```
0.8283778427831375
```

```
SST, RSS, R_squared = calculo_metricas(Y, predictions)
print("SST:", SST)
print("RSS:", RSS)
print("R^2:", R_squared)
```

```
SST: 2912.9348787590543
RSS: 914.0502713399505
R^2: 0.6862098504140446
```

```
beta0, beta1 = calculo_coef(X, Y)
print("Coeficiente beta1 (pendiente):", beta1)
print("Coeficiente beta0 (intercepto):", beta0)
print("Transformación inversa beta0:", 10**beta0 - 1)
```

```
Coeficiente beta1 (pendiente): 0.8082209032806918
Coeficiente beta0 (intercepto): 3.504703249458644
Transformación inversa beta0: 3195.7100724582033
```

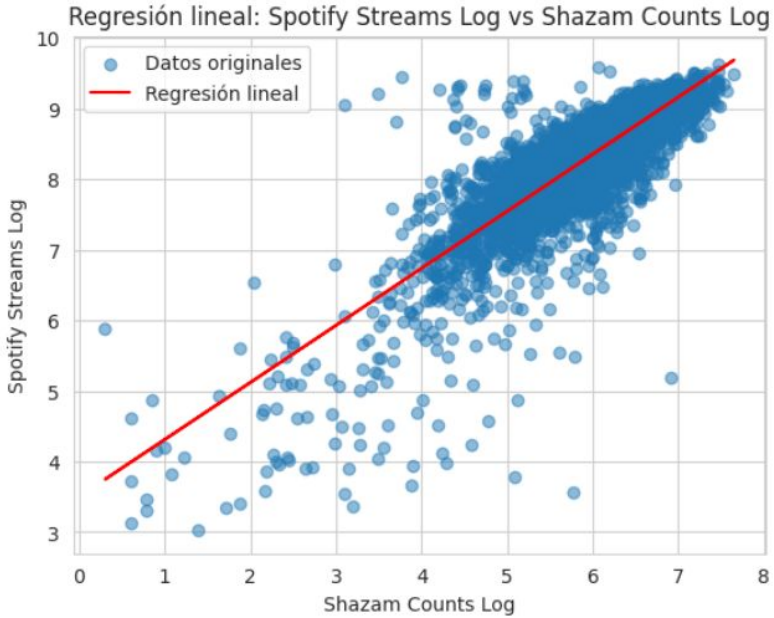
Regresión Lineal Simple - Transf. Log

OLS Regression Results

Dep. Variable:	Spotify Streams Log	R-squared:	0.686
Model:	OLS	Adj. R-squared:	0.686
Method:	Least Squares	F-statistic:	8682.
Date:	Sun, 06 Oct 2024	Prob (F-statistic):	0.00
Time:	01:35:36	Log-Likelihood:	-2718.3
No. Observations:	3972	AIC:	5441.
Df Residuals:	3970	BIC:	5453.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.5047	0.051	68.487	0.000	3.404	3.605
Shazam Counts Log	0.8082	0.009	93.176	0.000	0.791	0.825

Omnibus:	1364.381	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	19405.581
Skew:	-1.236	Prob(JB):	0.00
Kurtosis:	13.542	Cond. No.	40.8



Resultados

R-squared (R-cuadrado): 0.686 Interpretación: El valor de R-cuadrado indica qué porcentaje de la variación en la variable dependiente (Spotify Streams Log) puede explicarse por la variable independiente (Shazam Counts Log). En este caso, un valor de 0.686 significa que el 68.6% de la variación en el logaritmo de las reproducciones de Spotify puede explicarse por el logaritmo de las búsquedas en Shazam. Es un valor moderado, lo que sugiere que hay una relación importante entre ambas variables, aunque no es perfecta.

$$R^2 = \frac{SE}{ST}$$

Coeficiente de Shazam Counts: 0.808 Interpretación: Este coeficiente indica que, en promedio, por cada incremento de un 1% adicional de búsquedas en Shazam, se asocia aproximadamente un incremento en 0.808% de reproducciones adicionales en Spotify. Es decir, la relación es positiva, y un aumento en las búsquedas en Shazam lleva a un aumento en las reproducciones en Spotify.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

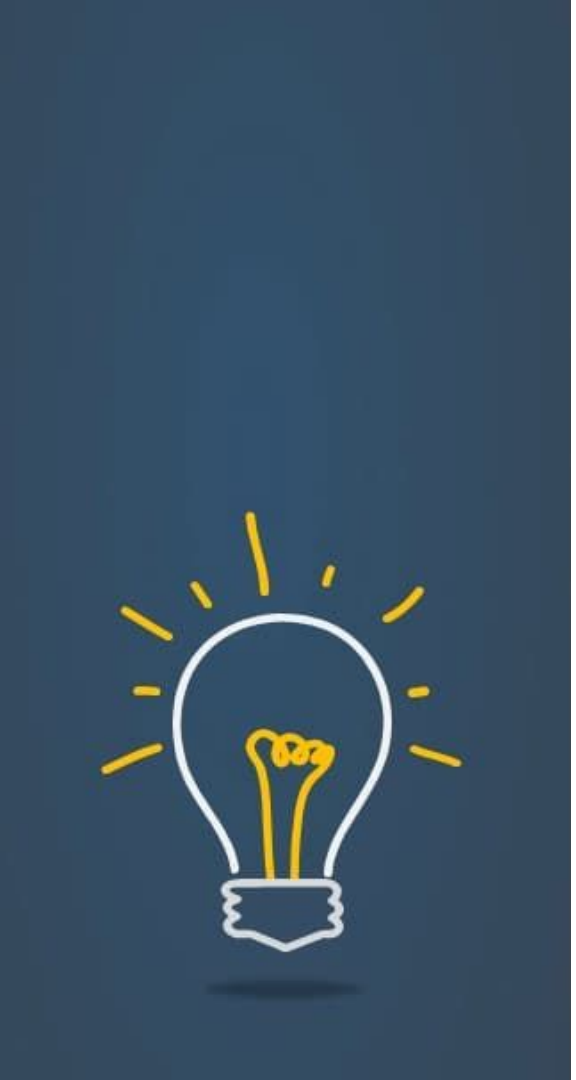
Resultados

Intercepto (const): 3.504 Interpretación: El valor del intercepto (3.504) representa el valor estimado del logaritmo de las reproducciones de Spotify cuando las búsquedas en Shazam son cero. Al invertir la transformación tenemos que el valor estimado de las reproducciones de Spotify son alrededor de 3196 cuando las búsquedas en Shazam son 0.

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}$$

P-valor de Shazam Counts: 0.000 Interpretación: El p-valor es prácticamente cero, lo que significa que la relación entre los logaritmos de las búsquedas en Shazam y los logaritmos de las reproducciones de Spotify es estadísticamente significativa. En otras palabras, es muy poco probable que esta relación ocurra por azar.

F-statistic: 8682 Interpretación: Un valor alto como este, junto con un p-valor asociado de 0.000, indica que el modelo en su conjunto es altamente significativo.



Conclusión

La regresión lineal simple muestra que existe una relación significativa y positiva entre los reconocimientos de canciones en en Shazam y las reproducciones en Spotify. Un mayor número de búsquedas en Shazam se asocia con un mayor número de reproducciones en Spotify.

Sin embargo, hay otros factores que explican el 46% restante de la variabilidad en las reproducciones, por lo que se podría necesitar incluir más variables para mejorar el ajuste del modelo.

Al aplicar una transformación logarítmica el modelo logra mejor la variabilidad de las reproducciones, aunque todavía queda un 31.4% de la variabilidad explicados por otros factores.