PREMIOS AL CINE



AYELÉN ALONSO

ENERO - 2024

OBJETIVOS

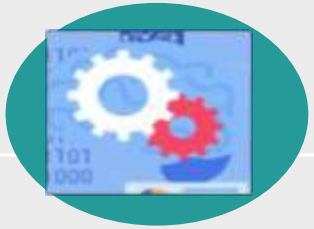
Comparar el comportamiento de la cantidad de premios recibidos por los distintos géneros de la industria cinematográfica.

Identificar si existe relación entre la recepción de premios y otras variables.

Analizar si hay posibilidades de que una película reciba o no un premio.

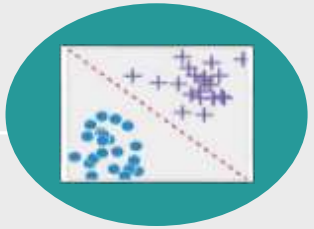
TAREAS REALIZADAS

EDA Y DATA WRANWLING



FEATURE ENGINEERING

REDUCCIÓN DE DIMENSIONALIDAD



ALGORITMOS DE CLASIFICACION

EDA Y DATA WRANWLING

ANÁLISIS PRELIMINAR DE DATOS

DICCIONARIO DE VARIABLES

```
print(df_peliculas.dtypes)
```

```
Título                          object
Género                          object
Año                           category
Director                        object
Duración                        object
Calificación                   float64
Ingresos (millones)            float64
Presupuesto (millones)         float64
País                            object
Premios                          int64
Espectadores                     int64
Críticas Positivas               int64
Críticas Negativas               int64
Popularidad                    float64
Recaudación en DVD (millones)  float64
Resultado (millones)           float64
Criticas                         int64
Premios_binario                  int64
genero_binario                   int64
genero_label                     int64
Categoria                     category
dtype: object
```

EDA Y DATA WRANWLING

ANÁLISIS PRELIMINAR DE DATOS

DETECCIÓN DE DUPLICADOS, NULOS Y ERRÓNEOS

```
# Detección y tratamiento de valores nulos

valores_nulos = df_peliculas.isnull().sum()
print(valores_nulos)
```

```
Título                          0
Género                          0
Año                             0
Director                        0
Duración                        0
Calificación                    0
Ingresos (millones)             0
Presupuesto (millones)          0
País                            0
Premios                         0
Espectadores                    0
Críticas Positivas              0
Críticas Negativas              0
Popularidad                     0
Recaudación en DVD (millones)   0
dtype: int64
```

No hay valores nulos, duplicados ni erróneos

```
# Detección y tratamiento de valores NaN

valores_NaN = df_peliculas.isna().sum()
print(valores_NaN)
```

```
Título                          0
Género                          0
Año                             0
Director                        0
Duración                        0
Calificación                    0
Ingresos (millones)             0
Presupuesto (millones)          0
País                            0
Premios                         0
Espectadores                    0
Críticas Positivas              0
Críticas Negativas              0
Popularidad                     0
Recaudación en DVD (millones)   0
dtype: int64
```

```
# Análisis datos duplicados

valores_duplicados = df_peliculas.duplicated().sum()
print (valores_duplicados)
```

```
0
```

EDA Y DATA WRANWLING

ANÁLISIS PRELIMINAR DE DATOS

ANÁLISIS ESTADÍSTICO PRELIMINAR

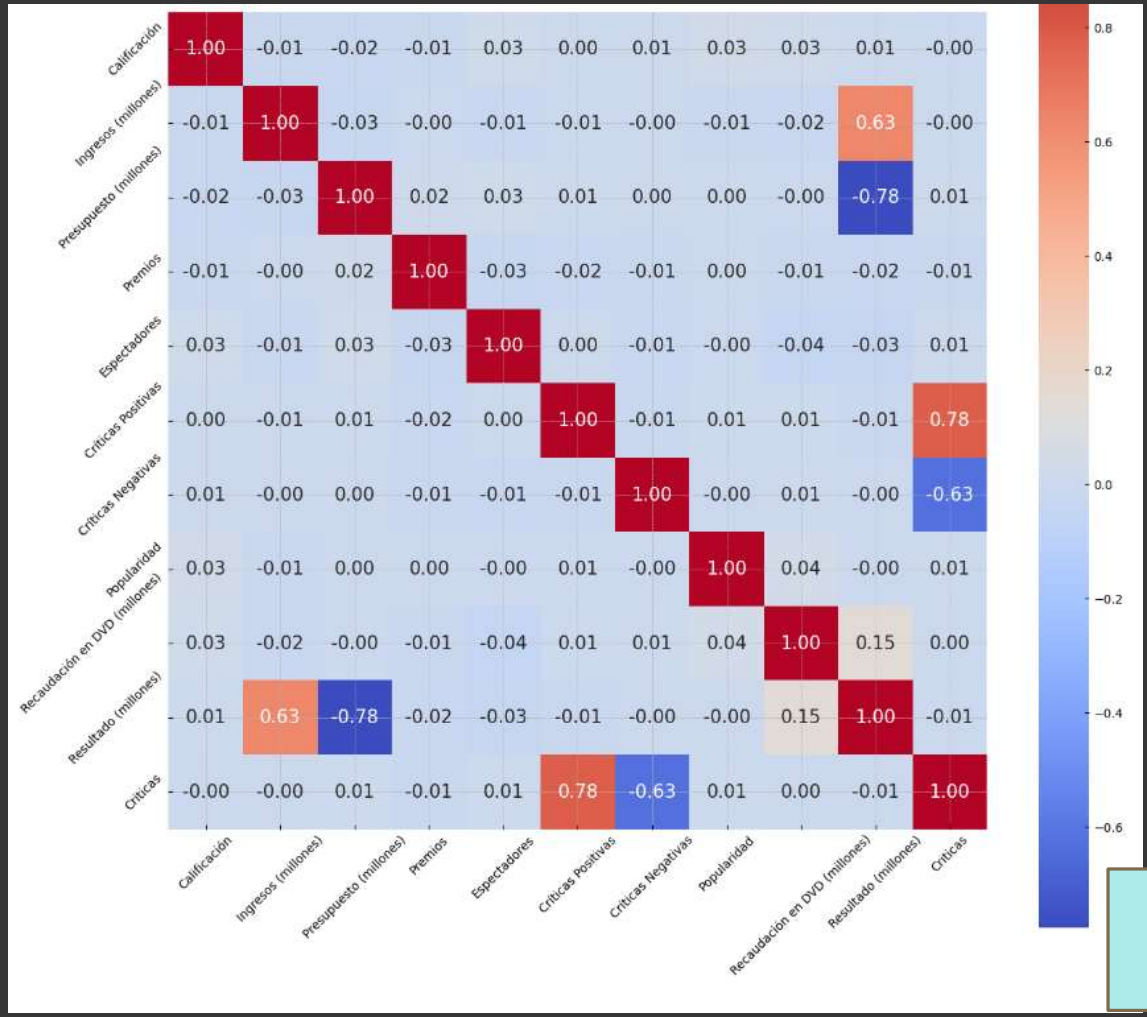
```
# Análisis estadístico preliminar

df_peliculas.describe()
```

	Calificación	Ingresos (millones)	Presupuesto (millones)	Premios	Espectadores	Críticas Positivas	Críticas Negativas	Popularidad	Recaudación en DVD (millones)	Resultado (millones)	Criticas
count	2000.000000	2000.00000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	4.899800	9.96965	17.357900	5.019000	511094.616500	75.359500	29.643500	4.980850	2.497050	-4.891200	45.716000
std	2.900113	5.86335	7.267939	3.159688	286494.162803	14.639447	11.659648	2.941201	1.460021	9.564332	18.781739
min	0.000000	0.00000	5.000000	0.000000	1123.000000	50.000000	10.000000	0.000000	0.000000	-28.000000	0.000000
25%	2.400000	4.90000	10.900000	2.000000	274333.500000	63.000000	19.000000	2.400000	1.200000	-11.800000	32.000000
50%	5.000000	9.90000	17.300000	5.000000	508741.500000	75.500000	29.000000	4.900000	2.500000	-5.100000	46.000000
75%	7.400000	15.10000	23.800000	8.000000	758798.250000	88.000000	40.000000	7.600000	3.800000	1.900000	60.000000
max	10.000000	20.00000	30.000000	10.000000	999530.000000	100.000000	50.000000	10.000000	5.000000	19.100000	89.000000

EDA Y DATA WRANWLING

ANÁLISIS PRELIMINAR DE DATOS

ANÁLISIS DE CORRELACIÓN ENTRE VARIABLES



Coeficiente de correlación de Pearson

EDA Y DATA WRANWLING

ANÁLISIS PRELIMINAR DE DATOS

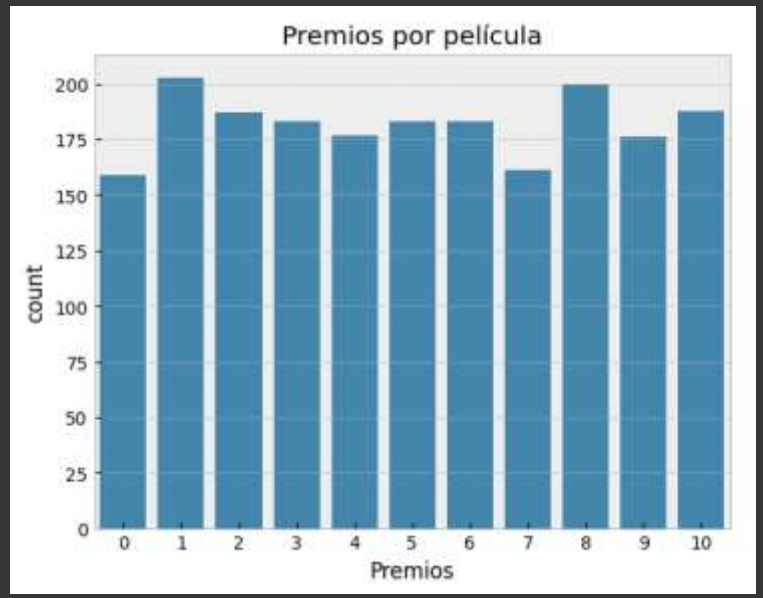
ANÁLISIS DE LA VARIABLE TARGET "PREMIOS"

MEDIDAS DE TENDENCIA CENTRAL

MEDIA	5,019
MEDIANA	5
MODA	1

DISTRIBUCIÓN DE LA VARIABLE

El test de Shapiro indica que los datos de la variable target no siguen una distribución normal

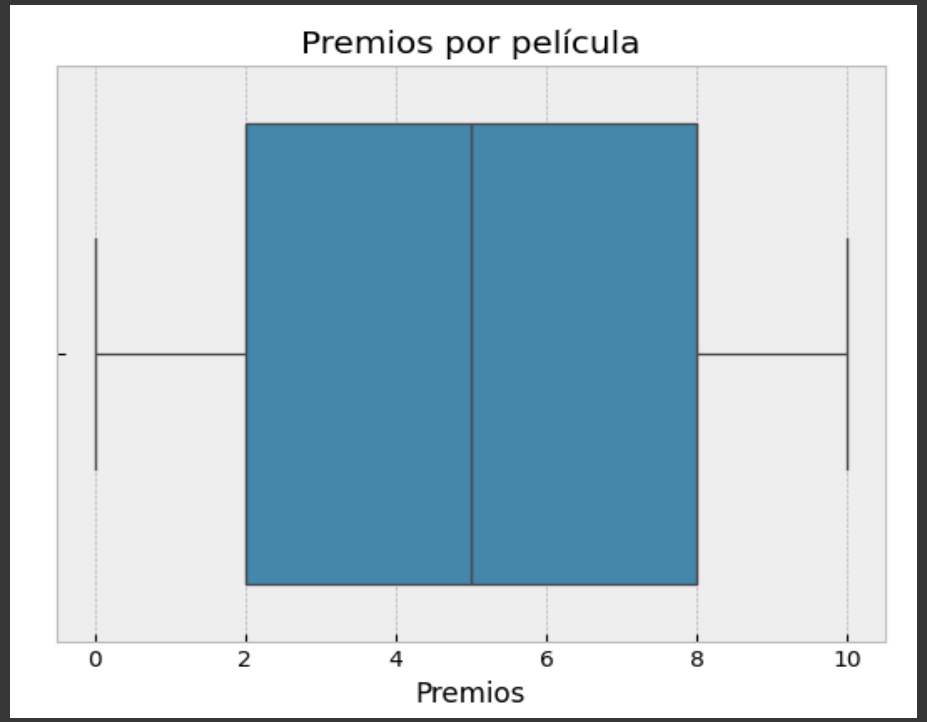


EDA Y DATA WRANWLING

ANÁLISIS PRELIMINAR DE DATOS

ANÁLISIS DE LA VARIABLE TARGET "PREMIOS"

DETECCIÓN DE OUTLIERS



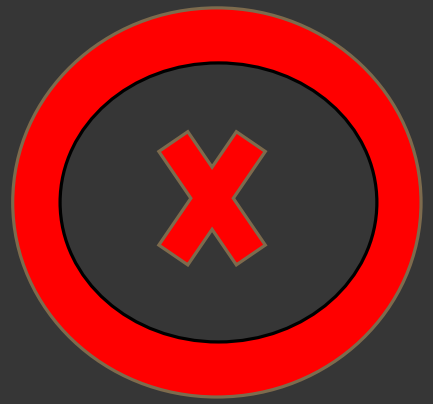
Gráficamente no se observan valores atípicos

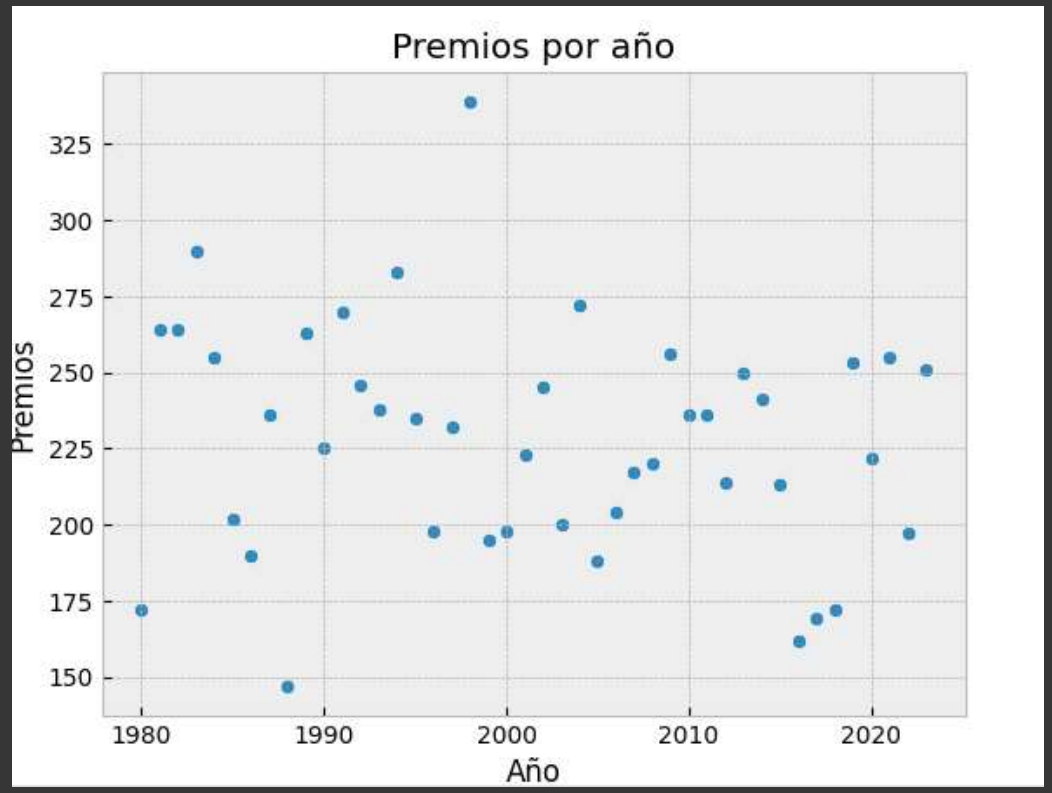
El cálculo del MAD arroja que no existen valores outliers en la variable target.

EDA Y DATA WRANWLING

ANÁLISIS DE HIPÓTESIS

¿A medida que pasan los años , más premios se entregan?





<u>Hipótesis nula:</u> No existe relación entre las variables año y premios

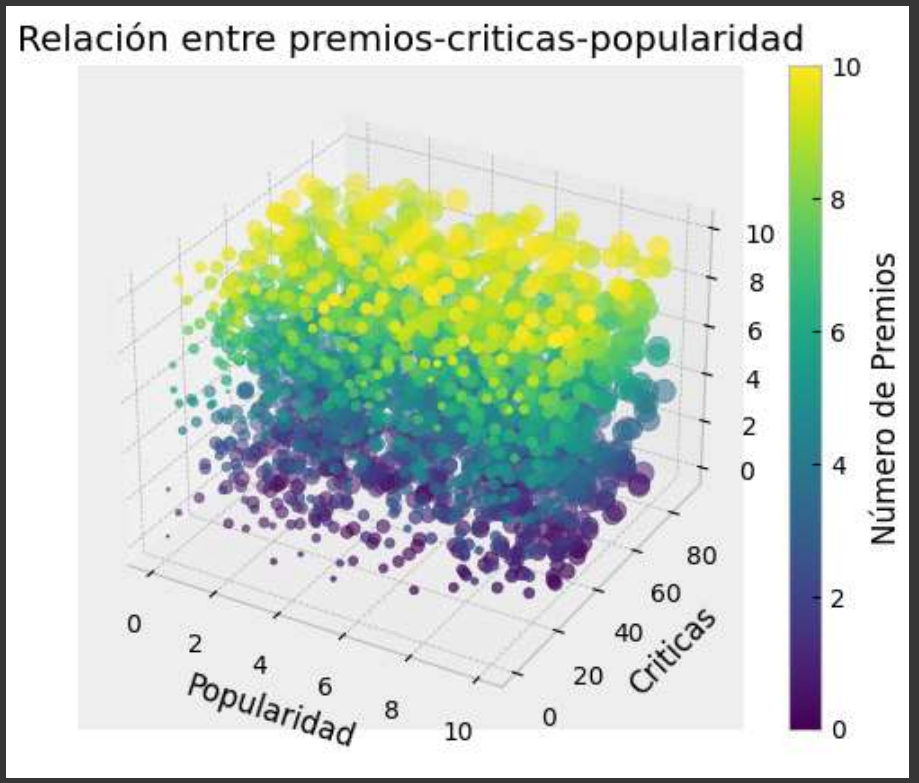
<u>Hipótesis alternativa:</u> Existe relación entre las variables año y premios

```
Coeficiente de Spearman: -0.18948259600000783
P-valor: 0.21798703796320967
```

EDA Y DATA WRANWLING

ANÁLISIS DE HIPÓTESIS

¿Existe relación entre la popularidad, las críticas obtenidas y la recepción de premios?



<u>Hipótesis nula:</u> No existe relación entre las variables premios, popularidad y críticas

<u>Hipótesis alternativa:</u> Existe relación entre las variables premios, popularidad y críticas

```
Coeficiente de Pearson: -0.007065761096766035
P-valor: 0.7521567217775721
```

Premios - Críticas

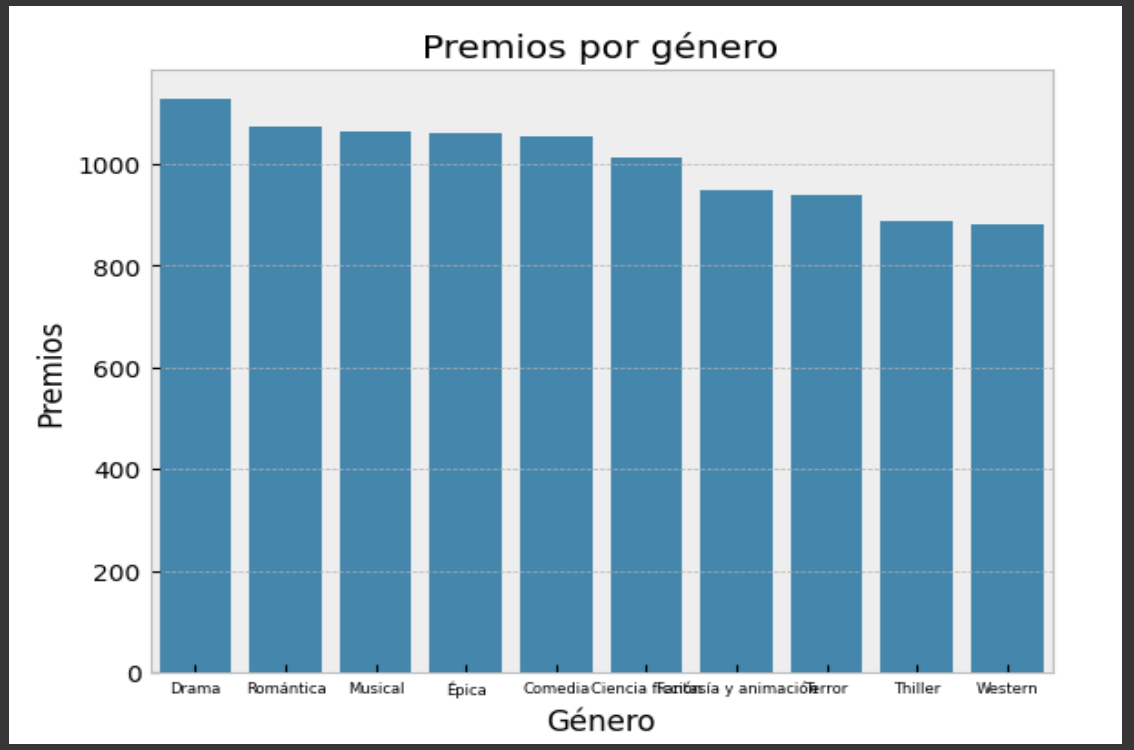
```
Coeficiente de Pearson: 0.0012503294066048198
P-valor: 0.9554362126679201
```

Premios - Popularidad

EDA Y DATA WRANWLING

ANÁLISIS DE HIPÓTESIS

El género dramático, ¿es el género más premiado?



<u>Hipótesis nula:</u> No existe relación entre las variables premios y género

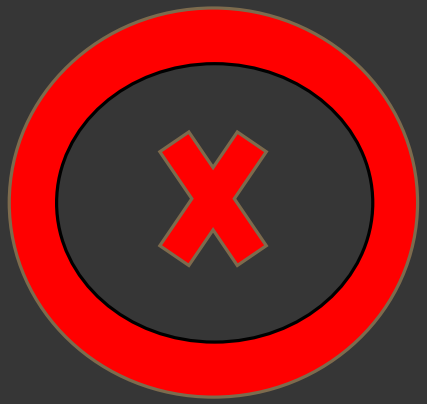
<u>Hipótesis alternativa:</u> Existe relación entre las variables premios y género

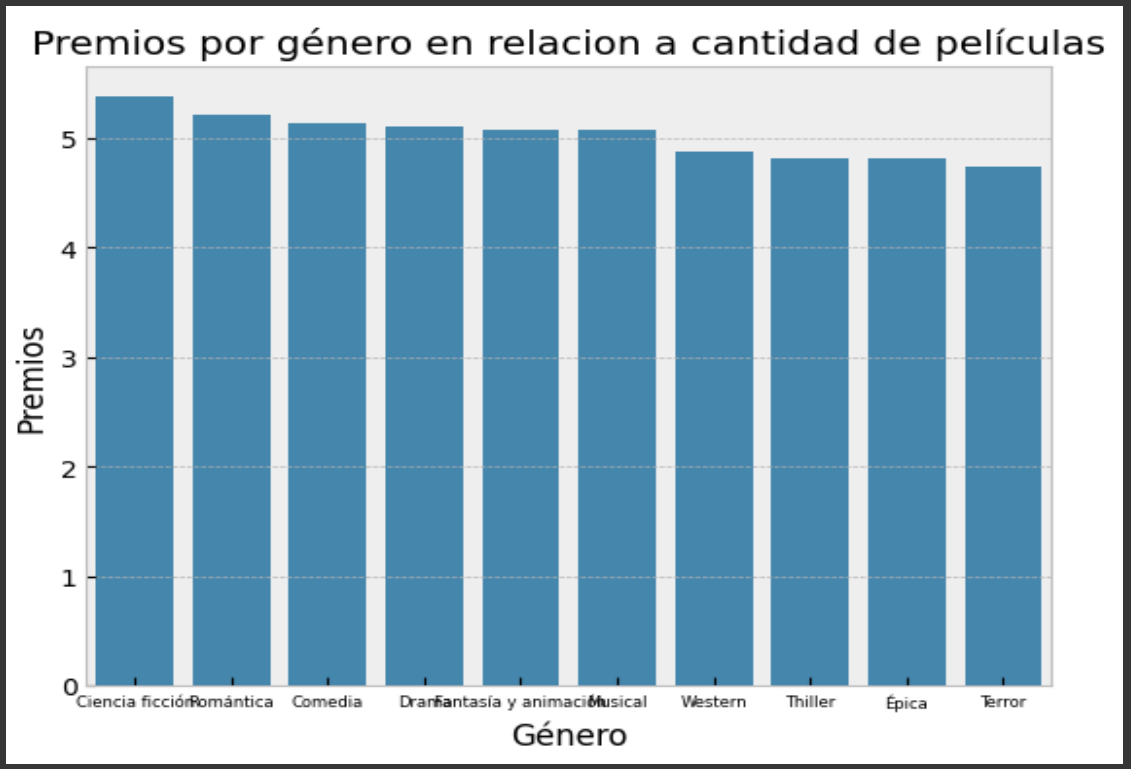
```
Coeficiente de Spearman: -0.35757575757575755
P-valor: 0.3103760917056799
```

EDA Y DATA WRANWLING

ANÁLISIS DE HIPÓTESIS

El género dramático, ¿es el género más premiado?





Hipótesis nula: No existe relación entre las variables premios y género

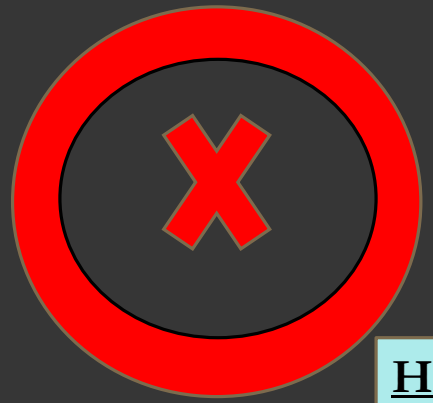
Hipótesis alternativa: Existe relación entre las variables premios y género

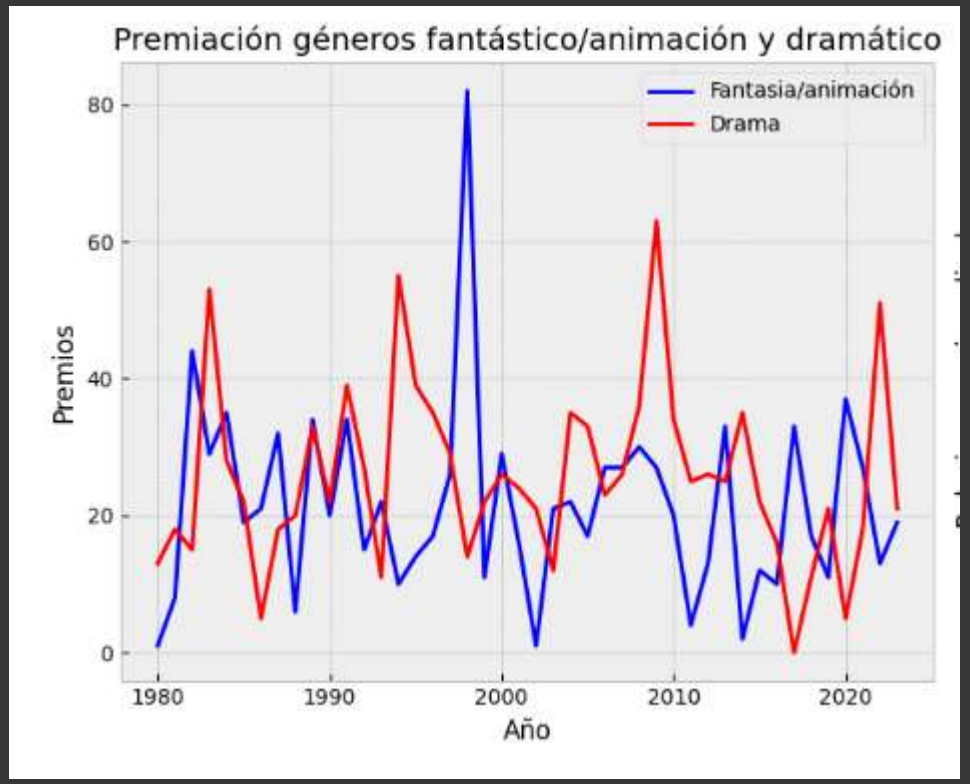
```
Coeficiente de Spearman: -0.7939393939393938
P-valor: 0.0060999233136969115
```

EDA Y DATA WRANWLING

ANÁLISIS DE HIPÓTESIS

¿Existe relación entre los premios recibidos por el género fantasía/animación en relación al género dramático a lo largo de la historia?






```
Coeficiente de Spearman: 0.1793279878979584
P-valor: 0.2441185101894924
```

Hipótesis nula: No existe relación entre las variables premios y género, para los géneros específicos drama y fantasía/animación

Hipótesis alternativa: Existe relación entre las variables premios y género, para los géneros específicos drama y fantasía/animación

FEATURE ENGINEERING

ENCODEO DE LAS VARIABLES

ENCODEO MANUAL VARIABLE "PREMIOS"

```
# Recuento

df_peliculas['Premios_binario'].value_counts(dropna=False)
```

```
1    1841
0     159
```

LABEL ENCODE VARIABLE "GÉNEROS"

```
# Recuento

df_peliculas['genero_label'].value_counts(dropna=False)
```

```
2    221
9    220
4    210
5    206
1    205
6    198
0    188
3    187
7    184
8    181
```

FEATURE ENGINEERING

FEATURE BINNING

CATEGORIZACIÓN DE LA VARIABLE “PREMIOS”

```
# Agrego columna para categorizar las peliculas segun los premios recibidos y convertir la columna a variable categórica
#"Muy Ganadoras": (9 <= valor <= 10)
#"Medianamente Ganadoras": (4 <= valor <= 8)
#"Poco Ganadoras": (0 <= valor <= 3)

cortes = [0, 3, 8, 10]
nombres =  ["Poco Ganadoras", 'Medianamente Ganadoras', 'Muy Ganadoras']
df_peliculas['Categoria'] = pd.cut(df_peliculas['Premios'], bins=cortes, labels=nombres)
value_counts = df_peliculas['Categoria'].value_counts().sort_values()
value_counts
```

```
Muy Ganadoras             364
Poco Ganadoras            573
Medianamente Ganadoras    904
Name: Categoria, dtype: int64
```

REDUCCIÓN DE DIMENSIONALI-DAD

PCA

```
print(lista_componentes)
print(modelo_pca.explained_variance_ratio_.round(2))
```

```
['PC1', 'PC2', 'PC3', 'PC4', 'PC5', 'PC6', 'PC7', 'PC8', 'PC9', 'PC10']
[0.2  0.18 0.11 0.1  0.1  0.1  0.09 0.09 0.02 0.  ]
```

No puede aplicarse PCA, ya que la mayoría de las variables numéricas se comportan como categóricas y la varianza acumulada no es significativa.

MCA

component	eigenvalue	% of variance	% of variance (cumulative)
0	0.007	0.65%	0.65%
1	0.006	0.58%	1.23%
2	0.006	0.58%	1.81%
3	0.006	0.57%	2.38%
4	0.006	0.56%	2.94%
5	0.006	0.55%	3.49%

No se logra una varianza acumulada significativa

MODELO DE CLASIFICACIÓN

APLICACIÓN DE ALGORITMOS: ANÁLISIS DE MÉTRICAS

VARIABLES INDEPENDIENTES: GÉNERO, POPULARIDAD, CRÍTICAS, CALIFICACIÓN Y ESPECTADORES

ALGORITMO / MÉTRICA		REGRESION LOGISTICA		KNN				ÁBOL DECISIÓN				SVM				LIGHT GBM	
		VALIDACION SIMPLE	SMOTE	VALIDACION SIMPLE	GRID SEARCH		RANDOMIZED SEARCH	VALIDACION SIMPLE	GRID SEARCH		RANDOMIZED SEARCH	VALIDACION SIMPLE	GRID SEARCH		RANDOMIZED SEARCH	HALVING GRID SEARCH	
					SIN SMOTE	CON SMOTE			SIN SMOTE	CON SMOTE			SIN SMOTE	CON SMOTE		SIN SMOTE	CON SMOTE
PRECISION	0 - No recibe premio	0,00	0,08	0,00	0,00	0,09	0,00	0,05	0,00	0,10	0,00	0,00	0,00	0,09	0,00	0,00	0,06
	1 - Recibe premio	0,92	0,92	0,92	0,92	0,93	0,92	0,92	0,92	0,94	0,92	0,92	0,92	0,93	0,92	0,92	0,92
RECALL	0 - No recibe premio	0,00	0,38	0,00	0,00	0,45	0,00	0,06	0,00	0,62	0,00	0,00	0,00	0,60	0,00	0,00	0,15
	1 - Recibe premio	1,00	0,61	1,00	1,00	0,59	1,00	0,90	1,00	0,51	1,00	1,00	1,00	0,46	1,00	1,00	0,81
F1-SCORE	0 - No recibe premio	0,00	0,13	0,00	0,00	0,14	0,00	0,06	0,00	0,17	0,00	0,00	0,00	0,15	0,00	0,00	0,09
	1 - Recibe premio	0,96	0,73	0,96	0,96	0,72	0,96	0,91	0,96	0,66	0,96	0,96	0,96	0,61	0,96	0,96	0,86
ACCURACY		0,92	0,59	0,92	0,92	0,58	0,92	0,83	0,92	0,52	0,92	0,92	0,92	0,47	0,92	0,92	0,76

ALGORITMO / MÉTRICA		RANDOM FOREST				XG BOOST	
		VALIDACION SIMPLE	GRID SEARCH		RANDOMIZED SEARCH	HALVING GRID SEARCH	
			SIN SMOTE	CON SMOTE		SIN SMOTE	CON SMOTE
PRECISION	0 - No recibe premio	0,00	0,00	**0,09**	0,00	0,00	0,08
	1 - Recibe premio	0,92	0,92	**0,93**	0,92	0,92	0,92
RECALL	0 - No recibe premio	0,00	0,00	**0,45**	0,00	0,00	0,19
	1 - Recibe premio	1,00	1,00	**0,62**	1,00	1,00	0,81
F1-SCORE	0 - No recibe premio	0,00	0,00	**0,15**	0,00	0,00	0,11
	1 - Recibe premio	0,96	0,96	**0,74**	0,96	0,96	0,86
ACCURACY		0,92	0,92	**0,61**	0,92	0,92	0,76

MEJORES ALTERNATIVAS

MODELO DE CLASIFICACIÓN

RANDOM FOREST

VARIABLES INDEPENDIENTES DEFINIDAS POR EL MODELO

Importancia de los predictores en el modelo

	predictores	importancia
4	Espectadores	0.104021
9	Resultado (millones)	0.095566
7	Popularidad	0.089314
8	Recaudación en DVD (millones)	0.088267
2	Ingresos (millones)	0.086476
10	Criticas	0.086089
0	Año	0.084102
1	Calificación	0.081218
3	Presupuesto (millones)	0.079629
6	Críticas Negativas	0.072619
5	Críticas Positivas	0.072402
12	genero_label	0.053920
11	genero_binario	0.006378

	precision	recall	f1-score	support
0	0.12	0.26	0.16	47
1	0.93	0.83	0.88	553
accuracy			0.79	600
macro avg	0.52	0.54	0.52	600
weighted avg	0.87	0.79	0.82	600

MÉTRICAS CON MEJORA DE HIPERPARÁMETROS, STRATIFIED K-FOLD Y APLICACIÓN DE SMOTE PARA BALANCEO

MODELO DE CLASIFICACIÓN

XG BOOST

VARIABLES INDEPENDIENTES DEFINIDAS POR EL MODELO

Importancia de los predictores en el modelo

	predictores	importancia
11	genero_label	0.11
3	Espectadores	0.10
7	Recaudación en DVD (millones)	0.10
9	Criticas	0.09
5	Críticas Negativas	0.09
4	Críticas Positivas	0.09
0	Calificación	0.08
8	Resultado (millones)	0.08
1	Ingresos (millones)	0.08
2	Presupuesto (millones)	0.08
6	Popularidad	0.07
10	genero_binario	0.03

	precision	recall	f1-score	support
0	0.08	0.13	0.10	47
1	0.92	0.88	0.90	553
accuracy			0.82	600
macro avg	0.50	0.50	0.50	600
weighted avg	0.86	0.82	0.84	600

MÉTRICAS CON MEJORA DE HIPERPARÁMETROS, STRATIFIED K-FOLD Y APLICACIÓN DE SMOTE PARA BALANCEO

CONCLUSIONES

LAS MEJORES MÉTRICAS EN EL MODELO DE CLASIFICACÓN SE OBTUVIERON CON LOS ALGORITMOS RANDOM FOREST Y XG BOOST, DEFINIENDO LAS VARIABLES INDEPENDIENTES SEGÚN SU IMPORTANCIA EN EL MODELO, DE ESTA FORMA CON LA MEJORA DE HIPERPARÁMETROS, LA ESTRATIFICACION Y EL BALANCEO DE CLASES SE LOGRA OBTENER UN BUEN RENDIMIENTO DE AMBOS MODELOS.