ANÁLISIS SPOTIFY

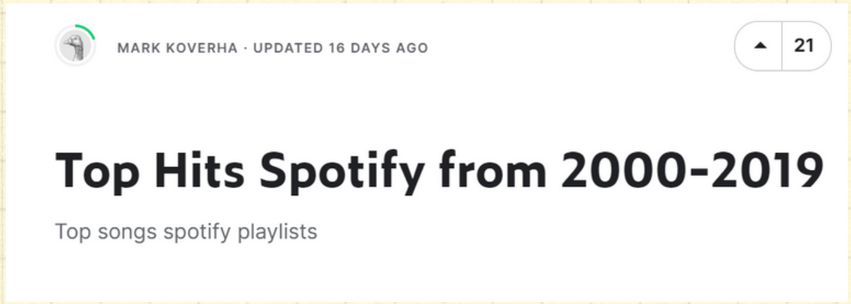
Iker Landeros - A01423214
Paulina Galindo - A01424818
Camila Turner - A01423579

Agenda de hoy

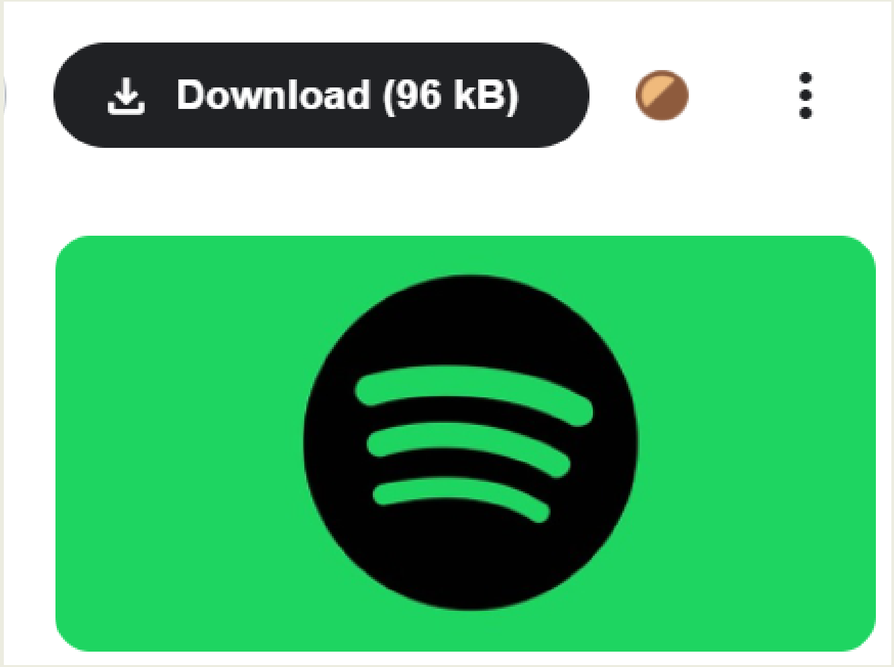
1. Base de datos
2. Preparación de datos
3. Modelo de regresión lineal
4. Análisis y conclusiones

Base de datos

Página: kaggle



https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019



Selección de los datos

DATOS A UTILIZAR:

- Artist
- ~~Song~~
- Year
- Popularity
- Genre
- Energy
- Danceabiliy

Selección de datos

```
artist = datos_spoti['artist']
year = datos_spoti['year']
popularity = datos_spoti['popularity']
genre = datos_spoti['genre']
energy  = datos_spoti['energy']
dance  = datos_spoti['danceability']
```

PROCESO EN PYTHON

```
spoti = {
    "artist": artist,
    "year": year,
    "popularity": popularity,
    "genre": genre,
    "energy": energy,
    "dance": dance
}
spoti = pd.DataFrame(spoti)
spoti
```

```
spotinum = {
    "popularity": popularity,
    "energy": energy,
    "dance": dance
}
spotinum = pd.DataFrame(spotinum)

ed = {
    "energy": energy,
    "dance": dance
}
ed = pd.DataFrame(ed)
```

Limpieza de datos

```
[4]  spoti.isnull().values.any() #Si imprime "false" es porque todos los valores son válidos

     dataset = spoti.dropna() # creamos un nuevo dataframe descartando los valores nulos o vacíos de nuestro dataframe datos_seleccionados

     dataset.isnull().sum() # validamos que no tenemos valores nulos en ninguna columna, todos deben dar cero

     # Como da 0 en cada columna, podemos avanzar.
```

```
artist        0
year          0
popularity    0
genre         0
energy        0
dance         0
dtype: int64
```

Preparación de datos

```
[ ]  dataset.columns
     x = dataset[['year']].values
     y = dataset['dance'].values
```

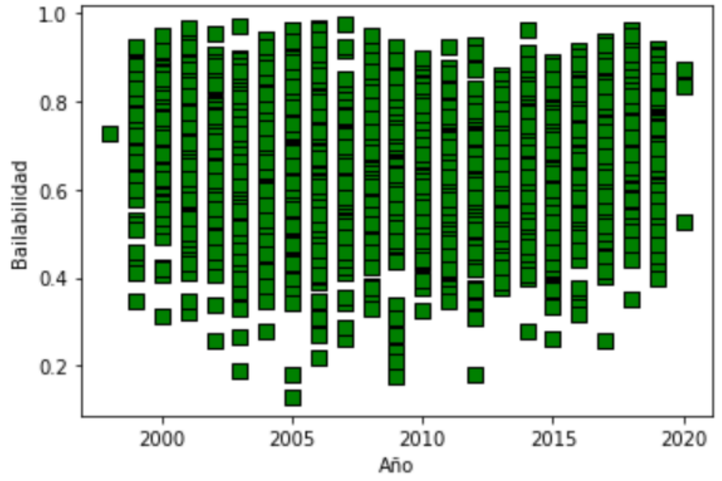
```
[ ]  x1 = dataset['dance']
     y1 = dataset[['energy']]
```

```
[ ]  x2 = dataset['genre']
     y2 = dataset[['popularity']]
```

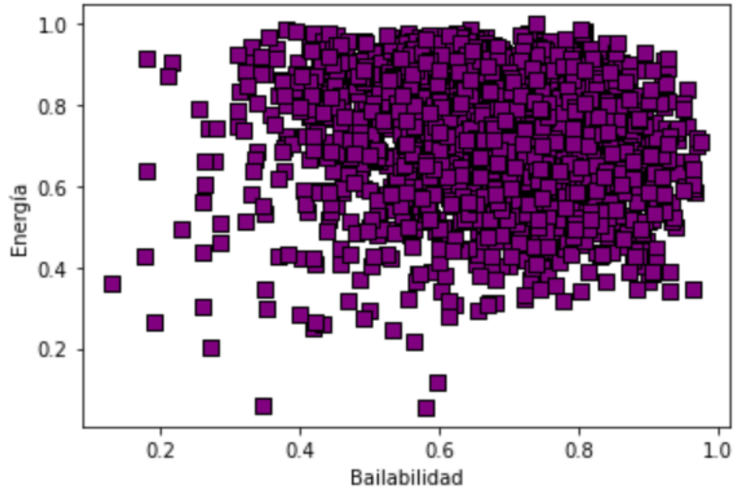
```
[ ]  from sklearn.model_selection import train_test_split # importamos la herramienta para dividir los datos de SciKit-Learn

     X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0) # asignación de los datos 80% para entrenamiento y 20%
```

Modelo de regresión lineal

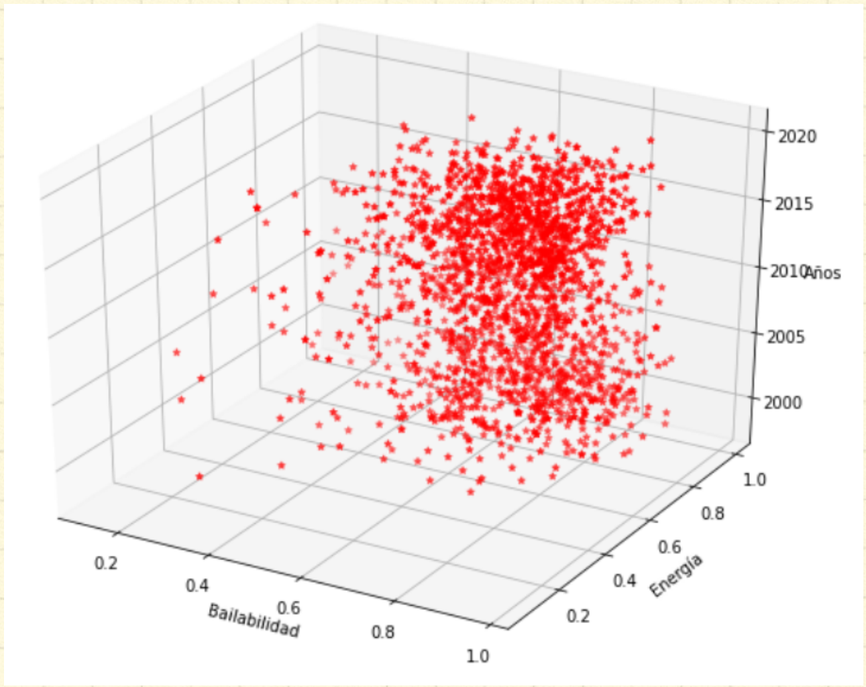
```
Coeficiente de correlación de Pearson:  0.03353246302645296
P-value:  0.1338473367206135
Text(0, 0.5, 'Bailabilidad')
```




```
Coeficiente de correlación de Pearson:  [-0.10403836408435745]
P-value:  3.1241303646488204e-06
Text(0, 0.5, 'Energía')
```



Modelo de regresión lineal múltiple



OLS Regression Results

Dep. Variable:	Y	R-squared:	0.012
Model:	OLS	Adj. R-squared:	0.011
Method:	Least Squares	F-statistic:	12.44
Date:	Fri, 13 May 2022	Prob (F-statistic):	4.28e-06
Time:	15:21:48	Log-Likelihood:	-6361.3
No. Observations:	2000	AIC:	1.273e+04
Df Residuals:	1997	BIC:	1.275e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2011.8060	0.931	2160.771	0.000	2009.980	2013.632
x_1	0.9378	0.933	1.005	0.315	-0.892	2.768
x_2	-4.0784	0.858	-4.754	0.000	-5.761	-2.396

Omnibus:	1447.458	Durbin-Watson:	0.226
Prob(Omnibus):	0.000	Jarque-Bera (JB):	120.538
Skew:	-0.096	Prob(JB):	6.69e-27
Kurtosis:	1.813	Cond. No.	14.3

Conclusiones

Después de realizar el análisis, nos dimos cuenta que las variables que seleccionamos no fueron las mejores. Los resultados que obtuvimos nos permitieron ver que las variables que escogimos no tenían mucha codependencia entre ellas.

Tal vez el modelo de regresión lineal no era la mejor opción o necesitábamos más variables.

***Lo podríamos trabajar en una segunda versión del proyecto*

Selección de los datos

- Artist
- ~~Song~~
- Year
- Popularity
- Genre
- Energy
- Danceability

DATOS A UTILIZAR:

- ~~Artist~~
- ~~Song~~
- ~~Year~~
- ~~Genre~~
- Popularity
- Energy
- Danceability
- + Instrumentalness
- + Speechiness

Selección de datos

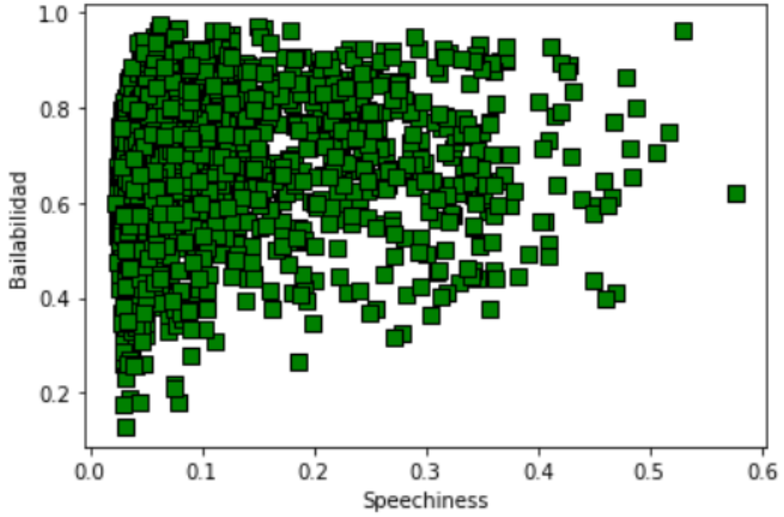
PROCESO EN PYTHON

```
artist = datos_spoti['artist']
year = datos_spoti['year']
popularity = datos_spoti['popularity']
genre = datos_spoti['genre']
energy  = datos_spoti['energy']
dance  = datos_spoti['danceability']
instrumental  = datos_spoti['instrumentalness']
speech  = datos_spoti['speechiness']
```

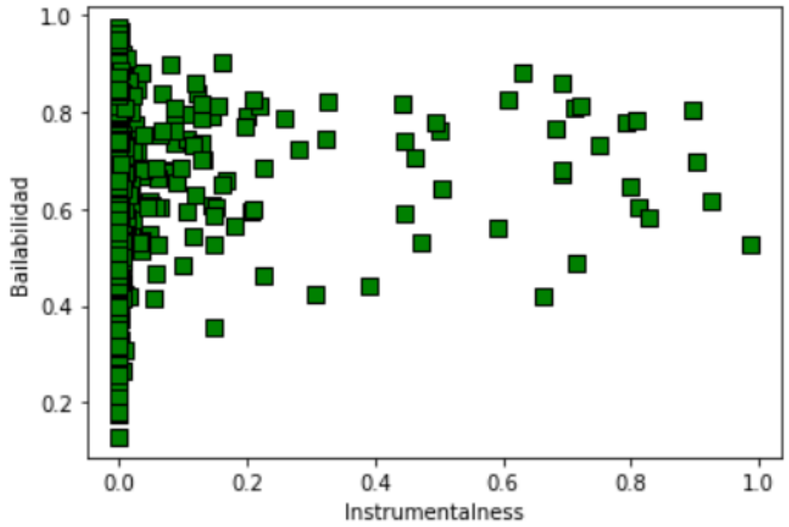
```
spoti = {
    "artist": artist,
    "year": year,
    "popularity": popularity,
    "genre": genre,
    "energy": energy,
    "dance": dance,
    "instrumental": instrumental,
    "speech": speech
}
spoti = pd.DataFrame(spoti)
spoti
```

Modelo de regresión lineal

```
Coeficiente de correlación de Pearson:  0.14558968510214182
P-value:  6.079013606698906e-11
Text(0, 0.5, 'Bailabilidad')
```

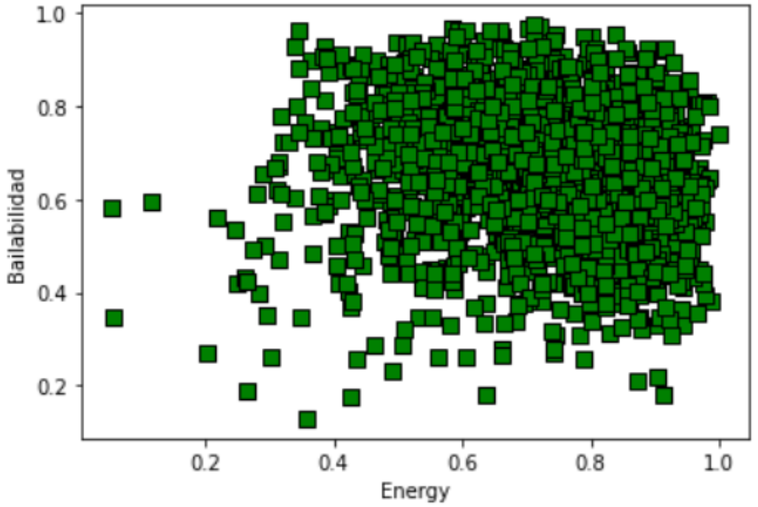



```
Coeficiente de correlación de Pearson:  0.023207307550550543
P-value:  0.2995712822605574
Text(0, 0.5, 'Bailabilidad')
```

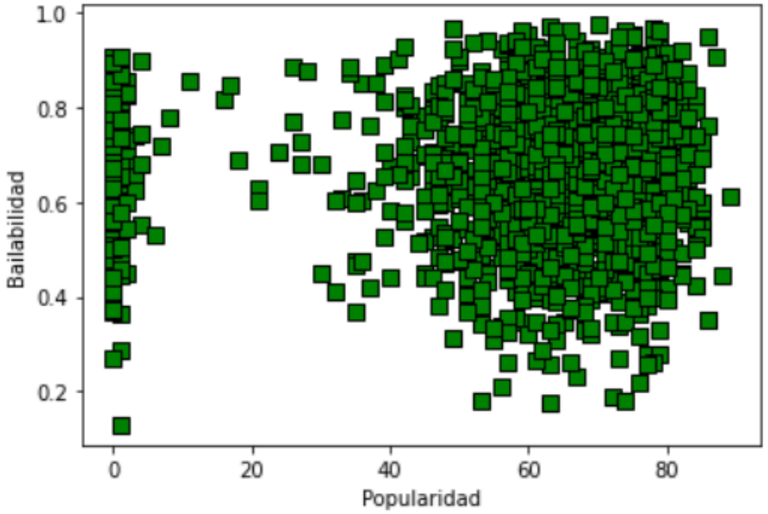


Modelo de regresión lineal

```
Coeficiente de correlación de Pearson:  -0.10403836408435757
P-value:  3.124130364648804e-06
Text(0, 0.5, 'Bailabilidad')
```




```
Coeficiente de correlación de Pearson:  -0.0035457302658119076
P-value:  0.8740854802420611
Text(0, 0.5, 'Bailabilidad')
```



Modelo de regresión lineal múltiple

OLS Regression Results

Dep. Variable:	Dy	R-squared:	0.032
Model:	OLS	Adj. R-squared:	0.030
Method:	Least Squares	F-statistic:	16.33
Date:	Fri, 13 May 2022	Prob (F-statistic):	3.59e-13
Time:	15:21:53	Log-Likelihood:	1121.1
No. Observations:	2000	AIC:	-2232.
Df Residuals:	1995	BIC:	-2204.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.7120	0.018	39.942	0.000	0.677	0.747
Vx[0]	0.2080	0.032	6.441	0.000	0.145	0.271
Vx[1]	-0.0895	0.020	-4.409	0.000	-0.129	-0.050
Vx[2]	0.0569	0.035	1.608	0.108	-0.012	0.126
Vx[3]	-4.091e-05	0.000	-0.282	0.778	-0.000	0.000

Omnibus:	85.701	Durbin-Watson:	1.823
Prob(Omnibus):	0.000	Jarque-Bera (JB):	95.868
Skew:	-0.522	Prob(JB):	1.52e-21
Kurtosis:	3.243	Cond. No.	734.

Variables independientes:

- Popularity
- Energy
- Instrumentalness
- Speechiness

Variable dependiente:

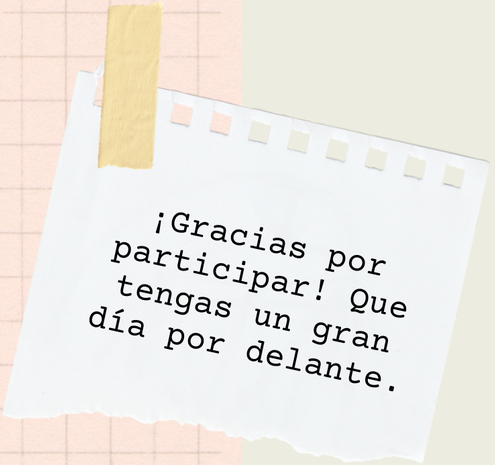
Danceability

Conclusiones 2.0

Después de realizar este segundo análisis, seguíamos sin obtener buenos resultados. El coeficiente de correlación seguía siendo muy bajo.

Lo más probable es que el modelo de regresión lineal no es la mejor opción para analizar las variables que seleccionamos y tal vez con otro modelo se puede llegar a obtener mejores resultados.

¡Gracias por tu atención!



Algunos links

Google colab:

https://colab.research.google.com/drive/1XMVV5Bpf-OMsWox5Xz8Cd-MB7_wADRPY?usp=sharing

Repositorio de Github:

https://github.com/A01423214/retoAnalitica.git