Proyecto Módulo 3: Procesamiento de Datos con Python

Carlos Álvarez Velasco

Analista de datos, BEDU

Agenda

- Problema.
- Objetivos.
- Limpieza de datos.
- Análisis de datos.
- Conclusiones.

Problema:

En la actualidad, el entretenimiento digital ha crecido de manera enorme en los últimos dos años, como son plataformas de streaming, redes sociales, videojuegos etc. Debido a esto, las críticas han sido más duras y la gente empieza a tener un mayor criterio.

Objetivos:

- Observar y manejar una base de datos desde cero, donde se requiere limpiar, ordenar y normalizar para su análisis apropiado para módulos posteriores.
- Aprender la librería de Pandas incluida en Python, para optimizar y limpiar una base de datos.
- De una base de datos relacionada a videojuegos, sacar conclusiones preliminares a base de un enfoque cualitativo.

Limpieza de datos

Se obtiene una base de datos de los juegos criticados por critica especializada (“score”, “critics”) y los usuarios o jugadores (“user_score”, “users”). Se extrajo del sitio “Kaggle”.

df

	name	platform	company	r-date	score	user_score	developer	genre	players	critics	users
0	The Legend of Zelda: Ocarina of Time	Nintendo64	Nintendo	11/23/1998	99	9.1	Nintendo	Action Adventure,Fantasy	1 Player	22	5749
1	Tony Hawk's Pro Skater 2	PlayStation	Sony	9/20/2000	98	7.4	NeversoftEntertainment	Sports,Alternative,Skateboarding	2-Jan	19	647
2	Grand Theft Auto IV	PlayStation3	Sony	4/29/2008	98	7.6	RockstarNorth	Action Adventure,Modern,Modern,Open-World	1 Player	64	3806
3	SoulCalibur	Dreamcast	SEGA	9/8/1999	98	8.5	Namco	Action,Fighting,3D	2-Jan	24	324
4	Grand Theft Auto IV	Xbox360	Microsoft	4/29/2008	98	7.9	RockstarNorth	Action Adventure,Modern,Modern,Open-World	1 Player	86	3364
...	...	...	...	...	...	...	...	...	...	...	...
17939	Vroom in the Night Sky	Switch	Nintendo	4/5/2017	17	3.1	Poisoft	Sports,Individual,Biking	No Online Multiplayer	15	105
17940	Leisure Suit Larry: Box Office Bust	PlayStation3	Sony	5/5/2009	17	1.9	Team17	Action Adventure,Adventure,Third-Person,Open-W...	No Online Multiplayer	11	45
17941	Yaris	Xbox360	Microsoft	10/10/2007	17	4.3	BackboneEntertainment	Driving,Racing,Arcade,Arcade,Automobile	2 Online	7	129
17942	Ride to Hell: Retribution	PC	Desktop	6/24/2013	16	1.3	Eutechnyx,DeepSilver	Driving,Modern,Racing,Motorcycle,Motocross,Mod...	No info	9	581
17943	Family Party: 30 Great Games Obstacle Arcade	WiiU	Nintendo	12/4/2012	11	2	ArtCo.,Ltd.	Miscellaneous,Party,Party,Party / Minigame	No Online Multiplayer	8	151

17944 rows × 11 columns

Limpieza de datos

- Se quitan los NAs.
- Se mete el año dentro del dataframe.
- Se cambian el tipo de dato de la calificación de usuarios a float, y la fecha de lanzamiento a datetime.
- Se hizo un Split a todos los géneros y desarrolladores. Se dejó la primera desarrolladora y el primer género.
- Se introdujo un nombre al index.
- Se limpiaron los datos de la columna “players”.

Limpieza de datos

	name	platform	company	release_date	score	user_score	players	critics	users	year	genre	developer
1	The Legend of Zelda: Ocarina of Time	Nintendo64	Nintendo	1998-11-23	99	9.1	1 Player	22	5749	1998	Action Adventure	Nintendo
2	Tony Hawk's Pro Skater 2	PlayStation	Sony	2000-09-20	98	7.4	2 players	19	647	2000	Sports	NeversoftEntertainment
3	Grand Theft Auto IV	PlayStation3	Sony	2008-04-29	98	7.6	1 Player	64	3806	2008	Action Adventure	RockstarNorth
4	SoulCalibur	Dreamcast	SEGA	1999-09-08	98	8.5	2 players	24	324	1999	Action	Namco
5	Grand Theft Auto IV	Xbox360	Microsoft	2008-04-29	98	7.9	1 Player	86	3364	2008	Action Adventure	RockstarNorth
...	...	...	...	...	...	...	...	...	...	...	...	...
17940	Vroom in the Night Sky	Switch	Nintendo	2017-04-05	17	3.1	No Online Multiplayer	15	105	2017	Sports	Poisoft
17941	Leisure Suit Larry: Box Office Bust	PlayStation3	Sony	2009-05-05	17	1.9	No Online Multiplayer	11	45	2009	Action Adventure	Team17
17942	Yaris	Xbox360	Microsoft	2007-10-10	17	4.3	2 Online	7	129	2007	Driving	BackboneEntertainment
17943	Ride to Hell: Retribution	PC	Desktop	2013-06-24	16	1.3	No info	9	581	2013	Driving	Eutechnyx
17944	Family Party: 30 Great Games Obstacle Arcade	WiiU	Nintendo	2012-12-04	11	2.0	No Online Multiplayer	8	151	2012	Miscellaneous	ArtCo.

17944 rows × 12 columns

```
] df_rewritten = df_3_renamed_3.copy()
  df_rewritten.index.names = ['game_id']
```

Limpieza de datos

```
[4]  def deleting_data():
         replacing = ['2-Jan', '3-Jan', '4-Jan', '5-Jan','6-Jan', '8-Jan', '10-Jan',
                      '12-Jan', '16-Jan', '24-Jan', 'Jan-64', 'Jan-32']
         return replacing
```

```
[5]  def adding_data():
         replacing = ['2 players', '3 players', '4 players', '5 players', '6 players',
                      '8 players', '10 players', '12 players', '16 players',
                      '24 players', '64 players', '32 players']
         return replacing
```

```
[6]  def reading_db():
         db = '/content/drive/MyDrive/colab_bedu/proyecto_bedu/games-data-exp.csv'
         return db
```

```
[7]  def writing_db():
         db = '/content/drive/MyDrive/colab_bedu/proyecto_bedu/games_rewritten.csv'
         return db
```

Análisis de datos

```
# Función para agrupar datos y obtener el resultado máximo de cada grupo.

def grouping_max(dataframe, groupby, agregation, criteria):
  grouping_max_data = dataframe.loc[dataframe.groupby(groupby)[agregation].idxmax()][criteria]
  return grouping_max_data
```

```
[50] # Función para normalizar una columna donde sus datos se repiten continuamente.

def one_row(dataframe, column, new_column_name,index_name):
  df_only_name = pd.DataFrame(dataframe[column], columns=[new_column_name])
  df_only_name.index.names = [index_name]
  df_only_name_no_duplicates = df_only_name.groupby(column).apply(list).reset_index()
  df_only_name_no_duplicates.index += 1
  df_only_name_no_duplicates = df_only_name_no_duplicates.drop(columns=[0])
  df_only_name_no_duplicates.index.names = [index_name]
  return df_only_name_no_duplicates
```

```
[51] # Función para filtrar datos que tengan más de una cierta cantidad de datos contados.

def filter_counter_more_than(dataframe,filter_more_than, column_name):
  df_original_count = dataframe[column_name].value_counts()
  df_count_more_than = df_original_count[df_original_count > filter_more_than]
  df_count_made = pd.DataFrame(df_count_more_than, columns=[column_name])
  df_count_made.index.names = [column_name]
  df_count_made = df_count_made.rename(columns={
    column_name: 'count'
  })
  return df_count_made
```

```
[ ] # Función para normalizar los datos en un dataframe

def normalize_column(original_df,name_column, destiny_dataframe):
  df_dict = original_df.to_dict("dict")
  df_pop = df_dict.pop(name_column)
  df_inverse = {v: k for k, v in df_pop.items()}
  df_unique = destiny_dataframe.replace(df_inverse)
  return df_unique
```

Análisis de datos

Se ordenan los 10 juegos con mayor número de críticas de la prensa y los jugadores.

```
[108] df_rewritten.sort_values('critics', ascending=False).head(10)
```

game_id	name	platform	company	release_date	score	user_score	players	critics	users	year	genre	developer
955	Final Fantasy VII Remake	PlayStation4	Sony	2020-04-10	87	8.1	No Online Multiplayer	126	6405	2020	Role-Playing	SquareEnix
2271	Ghost of Tsushima	PlayStation4	Sony	2020-07-17	83	9.2	No Online Multiplayer	122	17420	2020	Action Adventure	SuckerPunch
116	The Last of Us Part II	PlayStation4	Sony	2020-06-19	93	5.7	No Online Multiplayer	121	146262	2020	Action Adventure	NaughtyDog
70	God of War	PlayStation4	Sony	2018-04-20	94	9.2	No Online Multiplayer	118	16298	2018	Action Adventure	SCESantaMonica
1027	Marvel's Spider-Man	PlayStation4	Sony	2018-09-07	87	8.7	No Online Multiplayer	116	6051	2018	Action Adventure	InsomniacGames
690	Horizon Zero Dawn	PlayStation4	Sony	2017-02-28	89	8.4	No Online Multiplayer	115	9400	2017	Role-Playing	Guerrilla
4400	Paper Mario: The Origami King	Switch	Nintendo	2020-07-17	80	6.9	No Online Multiplayer	114	1462	2020	Role-Playing	IntelligentSystems
20	Super Mario Odyssey	Switch	Nintendo	2017-10-27	97	8.9	No Online Multiplayer	113	5546	2017	Action	Nintendo
164	Uncharted 4: A Thief's End	PlayStation4	Sony	2016-05-10	93	8.5	Up to 10	113	12333	2016	Action Adventure	NaughtyDog
1013	The Legend of Zelda: Link's Awakening	Switch	Nintendo	2019-09-20	87	8.4	No Online Multiplayer	111	1135	2019	Action Adventure	Nintendo

```
df_rewritten.sort_values('users', ascending=False).head(10)
```

game_id	name	platform	company	release_date	score	user_score	players	critics	users	year	genre	developer
116	The Last of Us Part II	PlayStation4	Sony	2020-06-19	93	5.7	No Online Multiplayer	121	146262	2020	Action Adventure	NaughtyDog
14809	Warcraft III: Reforged	PC	Desktop	2020-01-28	59	0.6	Online Multiplayer	46	30532	2020	Strategy	BlizzardEntertainment
111	The Witcher 3: Wild Hunt	PC	Desktop	2015-05-18	93	9.4	No Online Multiplayer	32	17537	2015	Action RPG	CDProjektRedStudio
2271	Ghost of Tsushima	PlayStation4	Sony	2020-07-17	83	9.2	No Online Multiplayer	122	17420	2020	Action Adventure	SuckerPunch
2952	Death Stranding	PlayStation4	Sony	2019-11-08	82	7.3	No Online Multiplayer	111	16949	2019	Action	KojimaProductions
70	God of War	PlayStation4	Sony	2018-04-20	94	9.2	No Online Multiplayer	118	16298	2018	Action Adventure	SCESantaMonica
13	The Legend of Zelda: Breath of the Wild	Switch	Nintendo	2017-03-03	97	8.6	No Online Multiplayer	109	15873	2017	Action Adventure	Nintendo
216	The Witcher 3: Wild Hunt	PlayStation4	Sony	2015-05-19	92	9.2	No Online Multiplayer	79	15749	2015	Action RPG	CDProjektRedStudio
59	The Last of Us Remastered	PlayStation4	Sony	2014-07-29	95	9.2	Up to 8	70	14563	2014	Action Adventure	NaughtyDog
16	Red Dead Redemption 2	PlayStation4	Sony	2018-10-26	97	8.4	Up to 32	99	14315	2018	Action Adventure	RockstarGames

Análisis de datos

Normalización de los datos

game_ID	name	platform	company	release_date	score	user_score	players	critics	users	year	genre	developer
1	The Legend of Zelda: Ocarina of Time	Nintendo64	2060	1998-11-23	99	9.1	1 Player	22	5749	1998	4	2060
2	Tony Hawk's Pro Skater 2	PlayStation	Sony	2000-09-20	98	7.4	2 players	19	647	2000	53	2029
3	Grand Theft Auto IV	PlayStation3	Sony	2008-04-29	98	7.6	1 Player	64	3806	2008	4	2557
4	SoulCalibur	Dreamcast	SEGA	1999-09-08	98	8.5	2 players	24	324	1999	3	2000
5	Grand Theft Auto IV	Xbox360	Microsoft	2008-04-29	98	7.9	1 Player	86	3364	2008	4	2557
...	...	...	...	...	...	...	...	...	...	...	...	...
17940	Vroom in the Night Sky	Switch	2060	2017-04-05	17	3.1	No Online Multiplayer	15	105	2017	53	2330
17941	Leisure Suit Larry: Box Office Bust	PlayStation3	Sony	2009-05-05	17	1.9	No Online Multiplayer	11	45	2009	4	3018
17942	Yaris	Xbox360	Microsoft	2007-10-10	17	4.3	2 Online	7	129	2007	16	281
17943	Ride to Hell: Retribution	PC	Desktop	2013-06-24	16	1.3	36	9	581	2013	16	992
17944	Family Party: 30 Great Games Obstacle Arcade	WiiU	2060	2012-12-04	11	2.0	No Online Multiplayer	8	151	2012	33	205

17944 rows × 12 columns

Análisis de datos

Resumen de juegos por género con más de 100 títulos en la lista.

genre	mean	median	std	count
Action	69.196937	71.0	13.000717	6007.0
Action Adventure	70.369193	72.0	12.803630	2454.0
Adventure	70.003693	72.0	10.535478	1083.0
Driving	69.436249	71.0	13.394666	949.0
Miscellaneous	69.860079	72.0	11.602400	1265.0
Puzzle	72.715116	74.0	8.910550	172.0
Racing	71.050980	71.0	10.740375	255.0
Role-Playing	72.502328	74.0	11.401128	1718.0
Simulation	68.589381	70.0	11.347566	565.0
Sports	72.474181	75.0	13.050383	1588.0
Strategy	71.463415	73.0	11.007314	1435.0

Análisis de datos

Resumen de juegos por consola con más de 10 títulos en la lista.

company	platform	mean	median	std	count
Desktop	PC	6.648824	7.10	1.822876	4592
Microsoft	Xbox	6.020050	7.30	2.984119	793
	Xbox360	6.582353	7.10	1.917227	1666
	XboxOne	5.849106	6.50	2.243014	1118
Nintendo	3DS	6.822750	7.40	1.724730	400
	DS	5.600959	7.20	3.201892	730
	GameBoyAdvance	5.809234	7.65	3.447370	444
	GameCube	6.838496	7.80	2.457474	452
	Nintendo64	8.080000	8.25	0.771710	70
	Switch	6.425490	7.30	2.477573	1122
	Wii	6.482229	7.30	2.403583	664
	WiiU	7.201075	7.50	1.306160	186
SEGA	Dreamcast	7.696000	8.00	1.587928	125
Sony	PSP	6.578363	7.40	2.469221	513
	PlayStation	6.671809	7.95	3.016586	188
	PlayStation2	7.019887	7.90	2.387408	1418
	PlayStation3	6.657210	7.10	1.754028	1269
	PlayStation4	6.145280	6.60	1.858742	1928
	PlayStationVita	7.143580	7.40	1.377089	257

Análisis de datos

Mejores juegos calificados según la crítica y usuarios para cada año.

game_id	name	score	year
1375	Full Throttle	86	1995
86	Sid Meier's Civilization II	94	1996
27	GoldenEye 007	96	1997
1	The Legend of Zelda: Ocarina of Time	99	1998
4	SoulCalibur	98	1999
2	Tony Hawk's Pro Skater 2	98	2000
14	Tony Hawk's Pro Skater 3	97	2001
18	Metroid Prime	97	2002
39	The Legend of Zelda: The Wind Waker	96	2003
23	Half-Life 2	96	2004
38	Resident Evil 4	96	2005
35	The Legend of Zelda: Twilight Princess	96	2006
6	Super Mario Galaxy	97	2007
3	Grand Theft Auto IV	98	2008
28	Uncharted 2: Among Thieves	96	2009
7	Super Mario Galaxy 2	97	2010
32	Batman: Arkham City	96	2011
167	Mass Effect 3	93	2012
10	Grand Theft Auto V	97	2013
9	Grand Theft Auto V	97	2014
24	Grand Theft Auto V	96	2015
164	Uncharted 4: A Thief's End	93	2016
13	The Legend of Zelda: Breath of the Wild	97	2017
16	Red Dead Redemption 2	97	2018
119	Divinity: Original Sin II - Definitive Edition	93	2019
58	Persona 5 Royal	95	2020

game_id	name	user_score	year
1375	Full Throttle	8.6	1995
110	Super Mario 64	9.2	1996
161	Castlevania: Symphony of the Night	9.2	1997
1922	Xenogears	9.2	1998
3163	Suikoden II	9.2	1999
6069	Resident Evil 2	9.2	2000
716	Shenmue II	9.2	2001
295	Resident Evil	9.2	2002
778	Warcraft III: The Frozen Throne	9.2	2003
1363	Tales of Symphonia	9.1	2004
29	Resident Evil 4	9.2	2005
8624	God Hand	9.2	2006
4566	GrimGrimoire	9.8	2007
246	Chrono Trigger	9.1	2008
279	Metroid Prime Trilogy	9.0	2009
13983	Metal Torrent	9.7	2010
2333	Ghost Trick: Phantom Detective	9.7	2011
249	Xenoblade Chronicles	9.2	2012
54	The Last of Us	9.2	2013
10266	Tengami	9.7	2014
111	The Witcher 3: Wild Hunt	9.4	2015
10609	Diaries of a Spaceport Janitor	9.7	2016
2958	The Evil Within 2	9.1	2017
70	God of War	9.2	2018
11683	Crystar	9.6	2019
3481	No More Heroes	9.6	2020

Conclusiones.

- A partir del procesamiento de datos con Pandas, se pudieron resolver problemas que tenía con SQL, como normalizar un gran conjunto de datos sin necesidad de Excel o VS.
- El uso de Google Colab o Jupyter Notebook es mejor que usar una IDE en ciertos casos, claro está, para manejar datos.
- El uso de funciones para reducir código es esencial para la óptima escritura de código.