ESCUELA POLITÉCNICA NACIONAL
ESCUELA DE FORMACIÓN DE TECNÓLOGOS
ANÁLISIS DE DATOS

Aplicación de análisis de datos

Autores:

- Eddy Morales
- Anthon Chang
- Paulo Cisneros

Fecha: lunes 3 de febrero del 2025

1. Definición del Caso de Estudio

Este caso de estudio analiza diversos datos provenientes de múltiples fuentes con el objetivo de obtener una visión integral de los diferentes factores que impactan en el área de interés. A través del cruce de información, se pretende identificar patrones, correlaciones y áreas de mejora, proporcionando una base sólida para la toma de decisiones y optimización de procesos.

2. Objetivos

2.1 Objetivo General

Analizar los datos provenientes de diversas fuentes para identificar patrones, correlaciones y áreas de mejora que contribuyan a optimizar procesos y tomar decisiones informadas.

2.2 Objetivos Específicos

1. Identificar las variables clave que afectan los procesos analizados.
2. Evaluar cómo las distintas fuentes de datos interactúan y contribuyen al análisis general.
3. Proponer acciones y estrategias basadas en los hallazgos obtenidos a partir de los datos analizados.

3. Descripción del Equipo de Trabajo y Actividades Realizadas

Equipo de Trabajo:

- Eddy Morales
- Anthon Chang
- Paulo Cisneros

Actividades:

1. Eddy Morales: Agreguen lo que hicieron
2. Anthon Chang: Limpieza de datos, la conexión al MySQL y al MongoDB y la creación de los dashboards.
3. Paulo Cisneros: Búsqueda de fuentes datos, limpieza de datos, Importación y exportación de datos en las bases de datos, creación de dashboards.

4. Cronograma de Actividades

Diagrama de Gantt:

Actividad	Lun 27	Mié 29	Jue 30	Vie 31	Sáb 1	Dom 2
Planificación del Proyecto						
Extraer Datasets						
Depuración de los Datos						
Exportar e Importar Datos						
Análisis de Sentimientos						
Realizar Dashboard						
Realización del Informe						
Presentación y Videos Expositivos						

5. Recursos y Herramientas Utilizadas

- **Recursos Humanos:**
  - Integrantes del grupo.
- **Herramientas Tecnológicas:**
  - Python para procesamiento de datos (Jupyter Notebook).
  - MongoDB como base de datos.
  - MySQL Workbench como base de datos.
  - SQL Server como base de datos.
  - CouchDB como base de datos.
  - Power BI para visualizaciones.

6. Arquitectura de la Solución

La arquitectura está hecha con las siguientes bases de datos SQL (MySQL Workbench,SQL Server) y NoSQL (MongoDB y CouchDB), el procesamiento de datos se hará en Python y visualización en Power BI.

Diagrama de Arquitectura

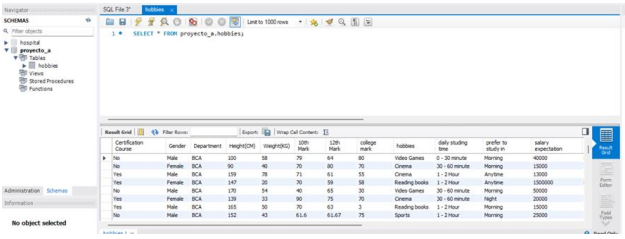
7. Extracción de Datos

Los datos fueron extraídos de la página web de “Kaggle”, que contiene data sets (.csv o .json) de la mayoría de los datos que encontramos. Y el tema de “Películas y series” fue extraído de una página web (https://www.rottentomatoes.com/m/joker_2019/reviews).

Luego se les paso a las bases de datos:

Hobbies

Datos en mysql:

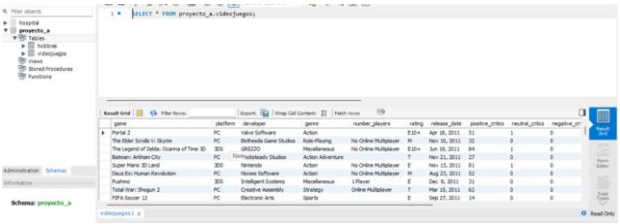


Datos en couchdb:

Gender	hobbies	daily studing	Department	part-time job
Male	Video Games	0 - 30 minute	BCA	No
Female	Cinema	30 - 60 minute	BCA	No
Male	Cinema	1 - 2 Hour	BCA	No
Female	Reading books	1 - 2 Hour	BCA	No
Male	Video Games	30 - 60 minute	BCA	No
Female	Cinema	30 - 60 minute	BCA	No
Male	Reading books	1 - 2 Hour	BCA	No
Male	Sports	1 - 2 Hour	BCA	No
Male	Video Games	0 - 30 minute	BCA	No
Male	Video Games	0 - 30 minute	BCA	No
Male	Video Games	1 - 2 Hour	BCA	No
Female	Cinema	2 - 3 hour	BCA	No
Female	Reading books	30 - 60 minute	BCA	No
Male	Cinema	0 - 30 minute	BCA	No

Videojuegos:

En mysql



En couch db:

Table | Metadata | {} JSON

game	developer	release_date	genre	metasc
Portal 2	Valve Software	Apr 18, 2011	Action	95
The Elder Scrolls V: Skyrim	Bethesda Game Studios	Nov 10, 2011	Role-Playing	94
The Legend of Zelda: Ocar…	GREZZO	Jun 19, 2011	Miscellaneous	94
Batman: Arkham City	Rocksteady Studios	Nov 21, 2011	Action Adventure	91
Super Mario 3D Land	Nintendo	Nov 13, 2011	Action	90
Deus Ex: Human Revolution	Nixxes Software	Aug 23, 2011	Action	90
Pushmo	Intelligent Systems	Dec 8, 2011	Miscellaneous	90
Total War: Shogun 2	Creative Assembly	Mar 15, 2011	Strategy	90
FIFA Soccer 12	Electronic Arts	Sep 27, 2011	Sports	89
Battlefield 3	EA DICE	Oct 25, 2011	Action	89
LIMBO	PLAYDEAD	Aug 2, 2011	Action	88
Assassin's Creed: Brother…	Ubisoft Montreal	Mar 22, 2011	Action Adventure	88
The Witcher 2: Assassins o…	CD Projekt Red Studio	May 17, 2011	Role-Playing	88
Dead Space 2	Visceral Games	Jan 25, 2011	Action	87

Noticias:

Table | Metadata | {} JSON | Create Document

_id	category	headline	date	short_descri…
211f58f8a1e9c1075af1eb…	ENTERTAINMENT	Justin Bieber And Lionel R…	2016-06-13	Hello? Goodbye.
211f58f8a1e9c1075af1eb…	THE WORLDPOST	Airstrike On Yemen Schoo…	2016-06-13	Aid group Doctors Without…
211f58f8a1e9c1075af1eb…	WOMEN	20 Reasons We're Terrible…	2016-06-13	Tonight, my husband and I…
211f58f8a1e9c1075af1eb…	HEALTHY LIVING	Diets, Doubts, And Dough…	2016-06-13	No, we are not, absolutely…
211f58f8a1e9c1075af1eb…	QUEER VOICES	Two Months After Pulse: A…	2016-06-13	It's sad to think that it coul…
211f58f8a1e9c1075af1eb…	SPORTS	The Olympics: More Rele…	2016-06-13	I have been obsessed wit…
211f58f8a1e9c1075af1eb…	IMPACT	Foster Care vs Adoption: …	2016-06-13	The Queensland Govern…
211f58f8a1e9c1075af1eb…	LATINO VOICES	From NPR's Latino USA: …	2016-06-13	Elizabeth Ramirez and he…
211f58f8a1e9c1075af1eb…	PARENTS	Why Ignoring My Son Wa…	2016-06-13	Today, like many of our da…
211f58f8a1e9c1075af1eb…	POLITICS	Why Donald Trump Is Sca…	2016-06-13	US presidential candidate …
211f58f8a1e9c1075af1eb…	SPORTS	Sole Russian Track And Fi…	2016-06-13	Darya Klishina is appealin…
211f58f8a1e9c1075af1eb…	SPORTS	Irish Rowers Deserve Gol…	2016-06-13	Brothers Paul and Gary O'…
211f58f8a1e9c1075af1eb…	PARENTS	Saying Goodbye: A Miscar…	2016-06-13	This has been a very emo…
211f58f8a1e9c1075af1eb…	SPORTS	Cowboys' Decal Proves R…	2016-06-13	Roger Goodell, commissio…

Mongodb :

Reseñas:

Type a query: { field: 'value' } or Generate query

Explain | Reset | Find

ADD DATA | EXPORT DATA | UPDATE | DELETE

25 | 1 - 25 of 514706

```
_id: ObjectId('679df7731a61c19b28bcc2541')
id_review : "R1245"
pre_review : "presentacion servicio exageradamente costoso informar ennla carta lis …"

_id: ObjectId('679df7731a61c19b28bcc2542')
id_review : "R1246"
pre_review : "presentacion plato calidad alimento vino experiencia recomendar"

_id: ObjectId('679df7731a61c19b28bcc2543')
id_review : "R1247"
pre_review : "reencontrame amigo querido peru visitar statera emocionar probar fabul…"

_id: ObjectId('679df7731a61c19b28bcc2544')
id_review : "R1248"
pre_review : "excelente experiencia comer maravilloso excelente insumo perfecto maxc…"

_id: ObjectId('679df7731a61c19b28bcc2545')
id_review : "R1249"
pre_review : "pasion viajar disfrutar grandeza pais sobretodo aprecio gastronomia pe…"

_id: ObjectId('679df7731a61c19b28bcc2546')
id_review : "R1250"
pre_review : "realmente fantastico disfrutar familia parejo reunion negocio perfecto…"
```

Liga española:

Type a query: { field: 'value' } or Generate query

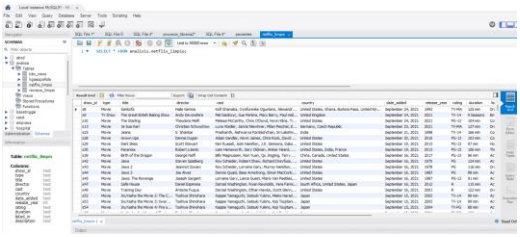
Explain | Reset | Find | Op

ADD DATA | EXPORT DATA | UPDATE | DELETE

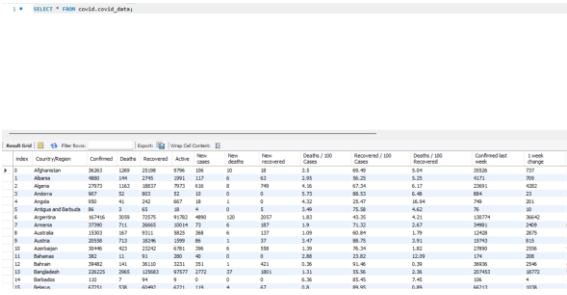
25 | 1 - 25 of 330

```
_id: ObjectId('679d7699a61c19b28bcc8f1a')
Indice : 1
Round : 33
Date : "2024-04-29 00:00:00"
Home Team : "Barcelona"
Away Team : "Valencia"
Home Score : 4
Away Score : 2
Total Goals : 6
Result : "Home"

_id: ObjectId('679d7699a61c19b28bcc8f1f')
Indice : 2
Round : 33
Date : "2024-04-28 00:00:00"
Home Team : "Betis"
Away Team : "Sevilla"
Home Score : 1
Away Score : 1
Total Goals : 2
Result : "Tie"
```

Mysql: Netflix



Mysql COVID:

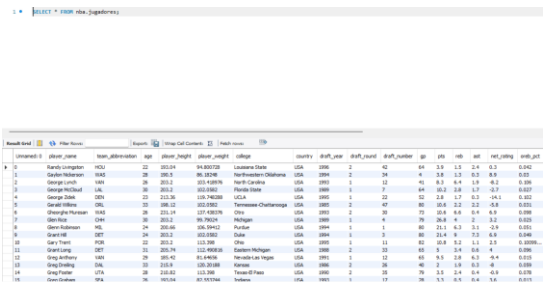


Mongo COVID:

```
_id: ObjectId('679da147f5bd746e0fb1c047')
index : 0
Country/Region : "Afghanistan"
Confirmed : 36263
Deaths : 1269
Recovered : 25198
Active : 9796
New cases : 106
New deaths : 10
New recovered : 18
Deaths / 100 Cases : 3.5
Recovered / 100 Cases : 69.49
Deaths / 100 Recovered : "5.04"
Confirmed last week : 35526
1 week change : 737
1 week % increase : 2.07
WHO Region : "Eastern Mediterranean"
```

```
_id: ObjectId('679da147f5bd746e0fb1c048')
index : 1
Country/Region : "Albania"
Confirmed : 4880
Deaths : 144
Recovered : 2745
Active : 1991
New cases : 117
New deaths : 6
New recovered : 63
```

Mysql NBA:



Mongo NBA:

```
_id: ObjectId('679db2a7f5bd746e0fb1c103')
Unnamed: 0 : 0
player_name : "Randy Livingston"
team_abbreviation : "HOU"
age : 22
player_height : 193.04
player_weight : 94.800728
college : "Louisiana State"
country : "USA"
draft_year : "1996"
draft_round : "2"
draft_number : "42"
gp : 64
pts : 3.9
reb : 1.5
ast : 2.4
net_rating : 0.3
oreb_pct : 0.042
dreb_pct : 0.071
usg_pct : 0.1689999999999999
ts_pct : 0.487
ast_pct : 0.248
season : "1996-97"
```

```
_id: ObjectId('679db2a7f5bd746e0fb1c104')
Unnamed: 0 : 1
player_name : "Gaylon Nickerson"
team_abbreviation : "WAS"
```

Mysql STARSBUCKS:

```
SELECT * FROM starbucks.tiendas;
```

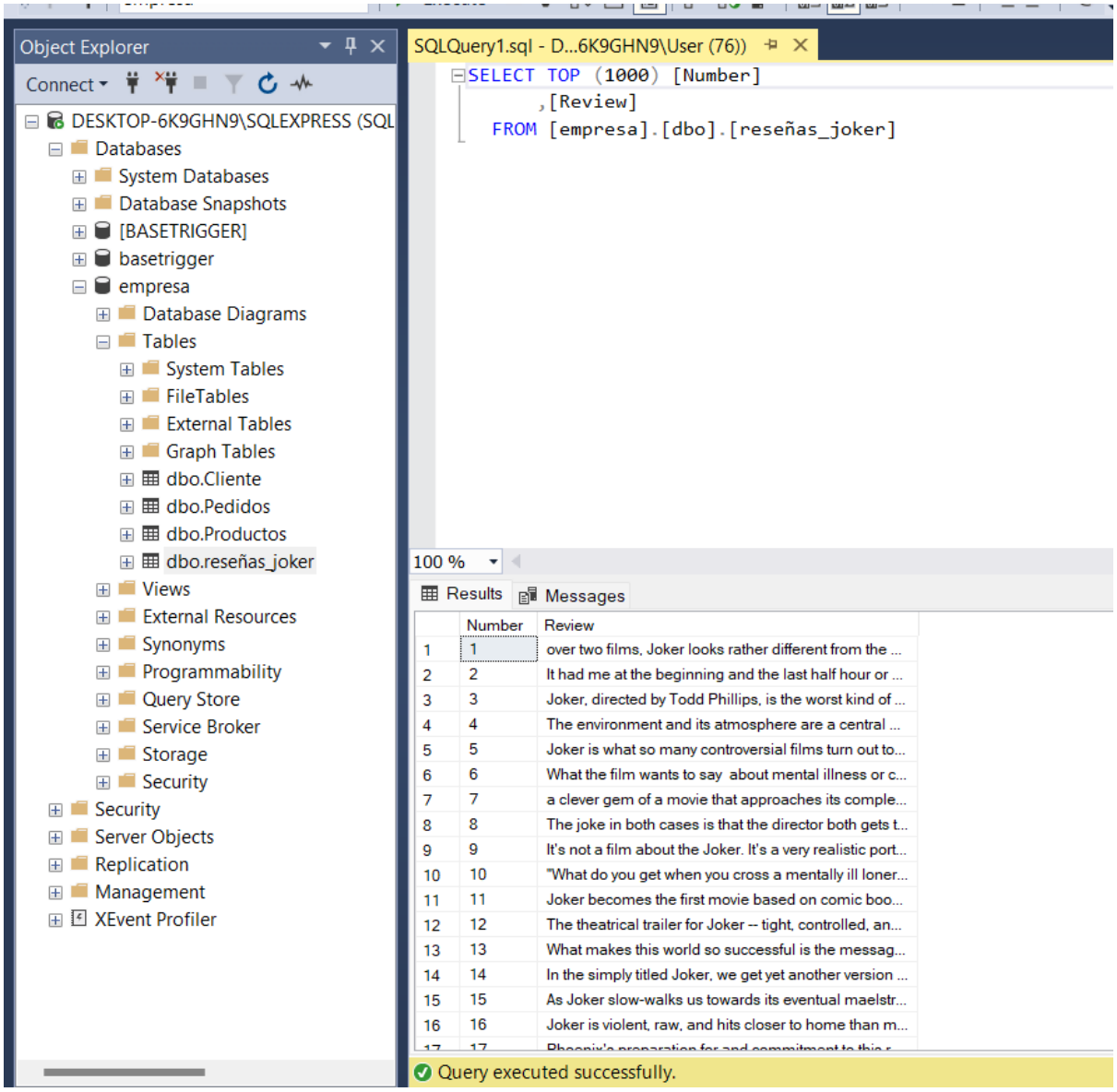
Brand	Store Number	Store Name	Ownership Type	Street Address	City	State/Province	Country	Postcode	Phone Number	Timezone	Longitude	Latitude
Starbucks	47370-257954	Meritxell, 96	Licensed	Av. Meritxell, 96	Andorra la Vella	7	AD	AD500	376818720	GMT+1:00 Europe/Andorra	1.53	42.51
Starbucks	22331-212325	Ajman Drive Thru	Licensed	1 Street 69, Al Jarf	Ajman	AJ	AE	00000	No disponible	GMT+04:00 Asia/Dubai	55.47	25.42
Starbucks	47559-264771	Dana Mall	Licensed	Sheikh Khalifa Bin Zayed St.	Ajman	AJ	AE	00000	No disponible	GMT+04:00 Asia/Dubai	55.47	25.39
Starbucks	22126-218029	Tourist Club Area	Licensed	Al Salam Street	Abu Dhabi	AZ	AE	00000	No disponible	GMT+04:00 Asia/Dubai	54.38	24.48
Starbucks	17127-178586	Al Ain Tower	Licensed	Khalifa Area, Abu Dhabi Island	Abu Dhabi	AZ	AE	00000	No disponible	GMT+04:00 Asia/Dubai	54.34	24.51
Starbucks	17088-162164	Dalma Mall, Ground Floor	Licensed	Dalma Mall, Mussafah	Abu Dhabi	AZ	AE	00000	No disponible	GMT+04:00 Asia/Dubai	54.49	24.4
Starbucks	18162-181165	Dalma Mall, Level 1	Licensed	Dalma Mall, Mussafah	Abu Dhabi	AZ	AE	00000	No disponible	GMT+04:00 Asia/Dubai	54.49	24.4
Starbucks	23339-237169	Deliverhero Yas Mall	Licensed	Yas Island	Abu Dhabi	AZ	AE	00000	No disponible	GMT+04:00 Asia/Dubai	54.61	24.48
Starbucks	30761-94622	Khalidiya Street	Licensed	Khalidiya St.	Abu Dhabi	AZ	AE	00000	2667050	GMT+04:00 Asia/Muscat	53.69	24.18
Starbucks	28423-203463	Eastern Mangroves	Licensed	Al Salam Street, The Mangroves	Abu Dhabi	AZ	AE	00000	No disponible	GMT+04:00 Asia/Dubai	54.38	24.48
Starbucks	20424-203468	Harbor Towers	Licensed	Corniche Street	Abu Dhabi	AZ	AE	00000	No disponible	GMT+04:00 Asia/Dubai	54.34	24.47
Starbucks	15379-122101	HCT Abu Dhabi Women's...	Licensed	Najda Street, Higher Colleges o...	Abu Dhabi	AZ	AE	31817	28408280	GMT+04:00 Asia/Dubai	54.37	24.48
Starbucks	32945-122105	Standard Chartered Bull...	Licensed	Khalidiya St., Beside Union Coo...	Abu Dhabi	AZ	AE	31817	26339275	GMT+04:00 Asia/Muscat	53.69	24.19
Starbucks	24181-139408	International Tower	Licensed	Capital Center, Adnec, Abu Dhabi	Abu Dhabi	AZ	AE	00000	No disponible	GMT+04:00 Asia/Dubai	54.7	24.3
Starbucks	24824-238307	Yas Mall 3	Licensed	YAS Island	Abu Dhabi	AZ	AE	00000	No disponible	GMT+04:00 Asia/Dubai	54.6	24.48
Starbucks	32228-118804	Blue Tower A	Licensed	Street No.4, Muroor Road, Gro...	Abu Dhabi	AZ	AE	00000	2649293	GMT+04:00 Asia/Muscat	55.74	24.21

Mongo STARSBUCKS:

```
_id: ObjectId('679db9c9f5bd746e0fb1f330')
Brand : "Starbucks"
Store Number : "47370-257954"
Store Name : "Meritxell, 96"
Ownership Type : "Licensed"
Street Address : "Av. Meritxell, 96"
City : "Andorra la Vella"
State/Province : "7"
Country : "AD"
Postcode : "AD500"
Phone Number : "376818720"
Timezone : "GMT+1:00 Europe/Andorra"
Longitude : "1.53"
Latitude : "42.51"
```

```
_id: ObjectId('679db9c9f5bd746e0fb1f331')
Brand : "Starbucks"
Store Number : "22331-212325"
Store Name : "Ajman Drive Thru"
Ownership Type : "Licensed"
Street Address : "1 Street 69, Al Jarf"
City : "Ajman"
State/Province : "AJ"
Country : "AE"
Postcode : "00000"
Phone Number : "No disponible"
Timezone : "GMT+04:00 Asia/Dubai"
Longitude : "55.47"
```

Sql server:



8. Análisis de Información

Se aplicaron técnicas de limpieza de datos para eliminar duplicados y valores nulos en cada uno de los dataframes.

A continuación, se adjuntan algunas capturas de pantalla de cómo se limpiaron cada uno de los dataframes.

Hobbies:

```python
#Revisar primeras filas
df_csv.head()
```

Python

	Certification Course	Gender	Department	Height(CM)	Weight(KG)	10th Mark	12th Mark	college mark	hobbies	daily studing time	prefer to study in	salary expectation	Do you like your degree?	willingness to pursue a career based on their degree	social medai & video	Tr
0	No	Male	BCA	100.0	58.0	79.0	64.0	80.0	Video Games	0 - 30 minute	Morning	40000	No	50%	1.30 - 2 hour	
1	No	Female	BCA	90.0	40.0	70.0	80.0	70.0	Cinema	30 - 60 minute	Morning	15000	Yes	75%	1 - 1.30 hour	
2	Yes	Male	BCA	159.0	78.0	71.0	61.0	55.0	Cinema	1 - 2 Hour	Anytime	13000	Yes	50%	More than 2 hour	
3	Yes	Female	BCA	147.0	20.0	70.0	59.0	58.0	Reading books	1 - 2 Hour	Anytime	1500000	No	50%	1.30 - 2 hour	
4	No	Male	BCA	170.0	54.0	40.0	65.0	30.0	Video Games	30 - 60 minute	Morning	50000	Yes	25%	1.30 - 2 hour	

```python
#Obtener informacion de los datos
df_csv.info()
```

Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 235 entries, 0 to 234
Data columns (total 19 columns):
 #   Column                                             Non-Null Count  Dtype
---  ------                                             --------------  -----
 0   Certification Course                               235 non-null    object
 1   Gender                                             235 non-null    object
 2   Department                                         235 non-null    object
 3   Height(CM)                                         235 non-null    float64
 4   Weight(KG)                                         235 non-null    float64
 5   10th Mark                                          235 non-null    float64
 6   12th Mark                                          235 non-null    float64
 7   college mark                                       235 non-null    float64
 8   hobbies                                            235 non-null    object
 9   daily studing time                                 235 non-null    object
 10  prefer to study in                                 235 non-null    object
 11  salary expectation                                 235 non-null    int64
 12  Do you like your degree?                           235 non-null    object
 13  willingness to pursue a career based on their degree 235 non-null  object
 14  social medai & video                               235 non-null    object
 15  Travelling Time                                    235 non-null    object
 16  Stress Level                                       235 non-null    object
 17  Financial Status                                   235 non-null    object
 18  part-time job                                      235 non-null    object
dtypes: float64(5), int64(1), object(13)
memory usage: 35.0+ KB
```

```python
#Eliminar filas duplicadas
df_csv = df_csv.drop_duplicates()


#Guardar en un archivo csv
df_csv.to_csv("C:/Users/Personal/OneDrive/Documentos/Archivos de programación/Jupyter/datos/hobbies_limpio.csv", index=False)
```

Archivo limpio: hobbies_limpio

Videojuegos:

Limpieza de datos archivo Videojuegos

```python
#Cargar el archivo
df2_csv = pd.read_csv("C:/Users/Personal/OneDrive/Documentos/Archivos de programación/Jupyter/datos/metacritic_games.csv")
```

[3] ✓ 0.0s Python

```python
#Revisar primeras filas
df2_csv.head()
```

[4] ✓ 0.0s Python

	game	platform	developer	genre	number_players	rating	release_date	positive_critics	neutral_critics	negative_critics	positive_users	neutral_users
0	Portal 2	PC	Valve Software	Action	NaN	E10+	Apr 18, 2011	51	1	0	1700	107
1	The Elder Scrolls V: Skyrim	PC	Bethesda Game Studios	Role-Playing	No Online Multiplayer	M	Nov 10, 2011	32	0	0	1616	322
2	The Legend of Zelda: Ocarina of Time 3D	3DS	GREZZO	Miscellaneous	No Online Multiplayer	E10+	Jun 19, 2011	84	1	0	283	26
3	Batman: Arkham City	PC	Rocksteady Studios	Action Adventure	NaN	T	Nov 21, 2011	27	0	0	240	34
	Super Mario				No Online		Nov 13					

```python
#Obtener informacion de los datos
df2_csv.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5699 entries, 0 to 5698
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   game              5699 non-null   object
 1   platform          5699 non-null   object
 2   developer         5685 non-null   object
 3   genre             5694 non-null   object
 4   number_players    3059 non-null   object
 5   rating            4433 non-null   object
 6   release_date      5699 non-null   object
 7   positive_critics  5699 non-null   int64
 8   neutral_critics   5699 non-null   int64
 9   negative_critics  5699 non-null   int64
 10  positive_users    5699 non-null   int64
 11  neutral_users     5699 non-null   int64
 12  negative_users    5699 non-null   int64
 13  metascore         5699 non-null   int64
 14  user_score        5699 non-null   int64
dtypes: int64(8), object(7)
memory usage: 668.0+ KB
```

```python
#Eliminar filas duplicadas
df2_csv = df2_csv.drop_duplicates()
```

✓ 0.0s

```python
#Guardar en un archivo csv
df2_csv.to_csv('C:/Users/Personal/OneDrive/Documentos/Archivos de programación/Jupyter/datos/videojuegos_limpio.csv', index=False)
```

Este es el archivo que quedo: videojuegos_limpio

Covid:

```python
covid.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Country/Region          187 non-null    object
 1   Confirmed               187 non-null    int64
 2   Deaths                  187 non-null    int64
 3   Recovered               187 non-null    int64
 4   Active                  187 non-null    int64
 5   New cases               187 non-null    int64
 6   New deaths              187 non-null    int64
 7   New recovered           187 non-null    int64
 8   Deaths / 100 Cases      187 non-null    float64
 9   Recovered / 100 Cases   187 non-null    float64
 10  Deaths / 100 Recovered  187 non-null    float64
 11  Confirmed last week     187 non-null    int64
 12  1 week change           187 non-null    int64
 13  1 week % increase       187 non-null    float64
 14  WHO Region              187 non-null    object
dtypes: float64(4), int64(9), object(2)
memory usage: 22.0+ KB
```

```
covid.isnull().sum()
```

```
Country/Region            0
Confirmed                 0
Deaths                    0
Recovered                 0
Active                    0
New cases                 0
New deaths                0
New recovered             0
Deaths / 100 Cases        0
Recovered / 100 Cases     0
Deaths / 100 Recovered    0
Confirmed last week       0
1 week change             0
1 week % increase         0
WHO Region                0
dtype: int64
```

Archivo limpio: COVID.csv

Nba:

```
nba_csv['college'].fillna('Otro', inplace=True)
```

```
nba_csv.to_csv('NBA.csv', index=False)
```

```
nba_csv.isnull().sum()
```

```
Unnamed: 0          0
player_name         0
team_abbreviation   0
age                 0
player_height       0
player_weight       0
college             0
country             0
draft_year          0
draft_round         0
draft_number        0
gp                  0
pts                 0
reb                 0
ast                 0
net_rating          0
oreb_pct            0
dreb_pct            0
usg_pct             0
ts_pct              0
ast_pct             0
season              0
dtype: int64
```

Archivo limpio: NBA.csv

Starbucks:

```
starbucks_csv['Street Address'].fillna('Desconocido', inplace=True)
```

```
starbucks_csv['City'].fillna('Ciudad Desconocido', inplace=True)
```

```
starbucks_csv['Postcode'].fillna('00000', inplace=True)
```

```
starbucks_csv['Phone Number'].fillna('No disponible', inplace=True)
```

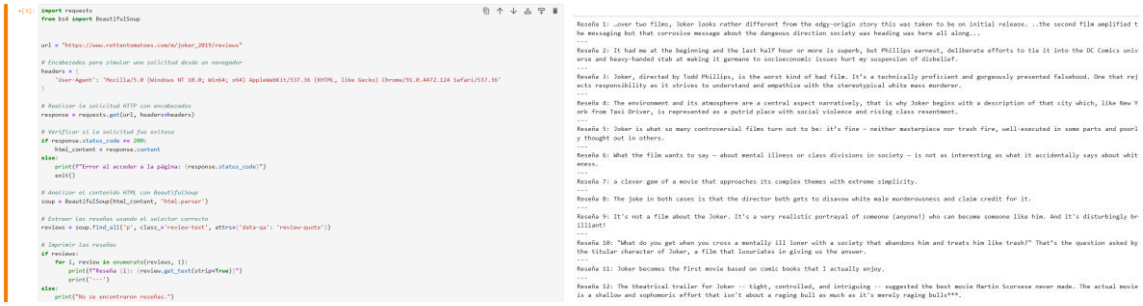
```
starbucks_csv['Longitude'].fillna('0.0', inplace=True)
```

```
starbucks_csv['Latitude'].fillna('0.0', inplace=True)
```

```
starbucks_csv.to_csv('STARBUCKS.csv', index=False)
```

Archivo limpio: STARBUCKS.csv

Web Scrapping: Reseñassobre la pelicula el Joker



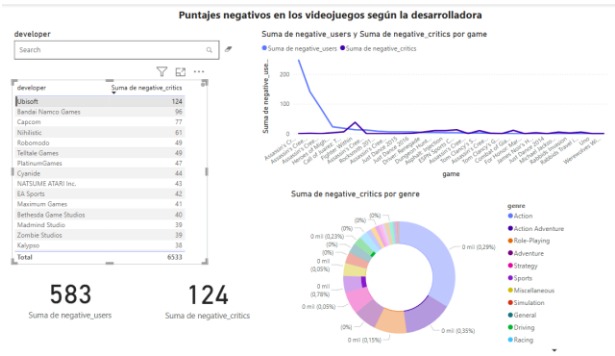
```
[3]: import requests
     from bs4 import BeautifulSoup


     url = "https://www.rottentomatoes.com/m/joker_2019/reviews"

     # Encabezados para simular una solicitud desde un navegador
     headers = {
         "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36"
     }

     # Realizar la solicitud HTTP con encabezados
     response = requests.get(url, headers=headers)

     # Verificar si la solicitud fue exitosa
     if response.status_code == 200:
         html_content = response.content
     else:
         print(f"Error al acceder a la página: {response.status_code}")
         exit()

     # Analizar el contenido HTML con BeautifulSoup
     soup = BeautifulSoup(html_content, 'html.parser')

     # Extraer las reseñas usando el selector correcto
     reviews = soup.find_all('p', class_='review-text', attrs={'data-qa': 'review-quote'})

     # Imprimir las reseñas
     if reviews:
         for i, review in enumerate(reviews, 1):
             print(f"Reseña {i}: {review.get_text(strip=True)}")
             print('---')
     else:
         print("No se encontraron reseñas.")
```

```
Reseña 1: ...over two films, Joker looks rather different from the edgy-origin story this was taken to be on initial release. ..the second film amplified t
he messaging but that corrosive message about the dangerous direction society was heading was here all along...
---
Reseña 2: It had me at the beginning and the last half hour or more is superb, but Phillips earnest, deliberate efforts to tie it into the DC Comics univ
erse and heavy-handed stab at making it germane to socioeconomic issues hurt my suspension of disbelief.
---
Reseña 3: Joker, directed by Todd Phillips, is the worst kind of bad film. It's a technically proficient and gorgeously presented falsehood. One that rej
ects responsibility as it strives to understand and empathize with the stereotypical white mass murderer.
---
Reseña 4: The environment and its atmosphere are a central aspect narratively, that is why Joker begins with a description of that city which, like New Y
ork from Taxi Driver, is represented as a putrid place with social violence and rising class resentment.
---
Reseña 5: Joker is what so many controversial films turn out to be: it's fine – neither masterpiece nor trash fire, well-executed in some parts and poorl
y thought out in others.
---
Reseña 6: What the film wants to say – about mental illness or class divisions in society – is not as interesting as what it accidentally says about whit
eness.
---
Reseña 7: a clever gem of a movie that approaches its complex themes with extreme simplicity.
---
Reseña 8: The joke in both cases is that the director both gets to disavow white male murderousness and claim credit for it.
---
Reseña 9: It's not a film about the Joker. It's a very realistic portrayal of someone (anyone!) who can become someone like him. And it's disturbingly br
illiant!
---
Reseña 10: "What do you get when you cross a mentally ill loner with a society that abandons him and treats him like trash?" That's the question asked by
the titular character of Joker, a film that luxuriates in giving us the answer.
---
Reseña 11: Joker becomes the first movie based on comic books that I actually enjoy.
---
Reseña 12: The theatrical trailer for Joker -- tight, controlled, and intriguing -- suggested the best movie Martin Scorsese never made. The actual movie
is a shallow and sophomoric effort that isn't about a raging bull as much as it's merely raging bulls***.
```

9. Visualización de Información

Videojuegos

- ¿Cuál es la desarrolladora que más críticas negativas tuvo y cuál es su juego más criticado?

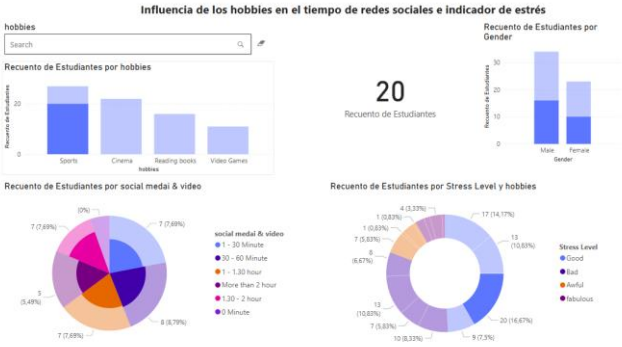


- ¿Cuál es la desarrolladora que más críticas positivas tuvo más gustado?

Hobbies y actividades

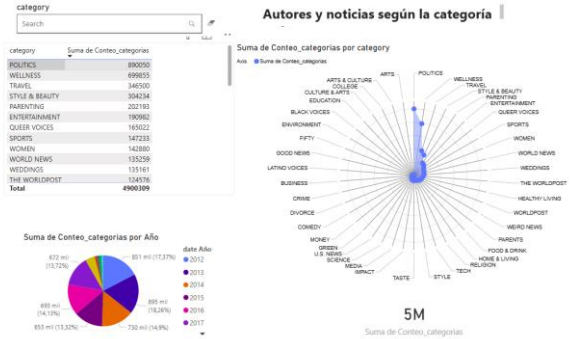
- ¿Cuál es un hobby que se puede practicar y que permite tener tiempo para estudiar y trabajar a medio tiempo?

- ¿Cómo influyen los hobbies que practicamos en la reducción del estrés y cómo pueden servir como indicadores de bienestar?



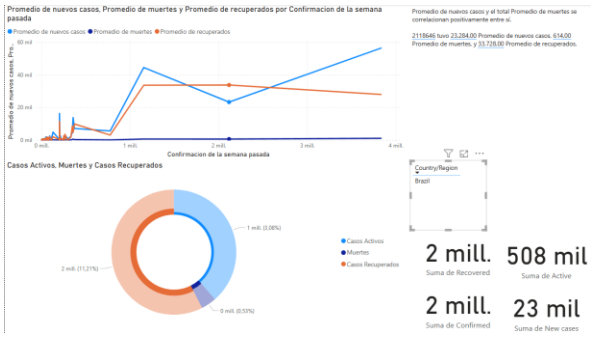
Noticias mundiales

- ¿Cuáles son las tendencias mundiales en las noticias?



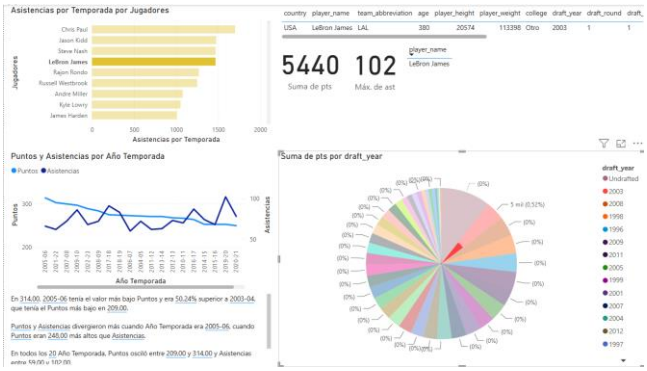
Covid

- ¿Cuál es la cantidad de casos confirmados, muertes y nuevos casos de COVID-19 en Brasil?

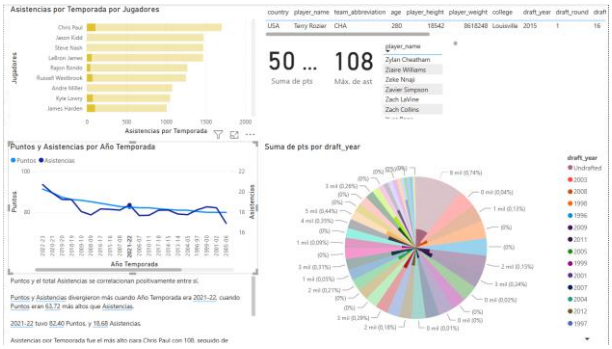


NBA

- ¿Cuál ha sido el mejor año de Lebrón James en términos de puntos y asistencias?

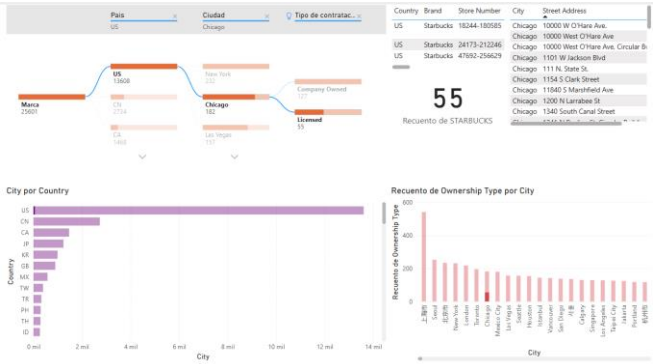


- ¿Cuál fue la suma total de puntos de los jugadores en la temporada 2021?



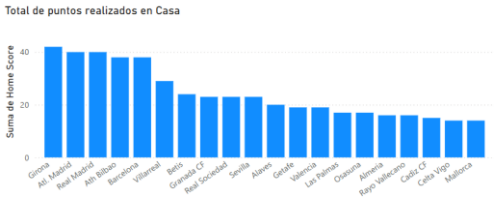
Starbucks

- ¿Cuál es el número total de locales de Starbucks con licencia en Chicago?

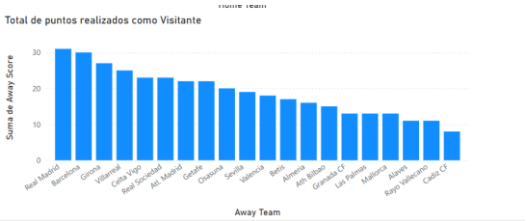


Liga española:

¿Cuáles son los puntos obtenidos por los equipos en Casa?

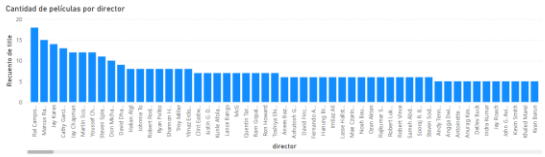


¿Cuáles son los puntos obtenidos por los equipos de visitante?

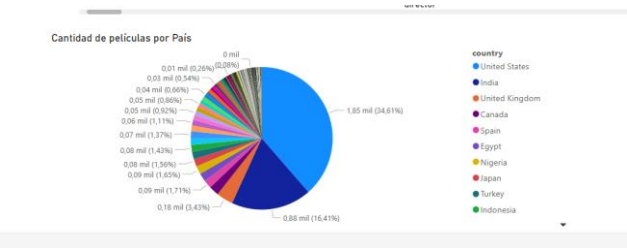


Netflix:

¿Cuáles son los tres directores con mayor cantidad de producciones realizadas?



¿Cuáles son los tres países con más producciones realizadas?



Reseña Comida Peruana:

¿Cuáles son las 5 palabras más frecuentes?:

Reseña película el Joker:

¿Cuáles son las palabras más frecuentes?:

10. Resultados Obtenidos

Videojuegos

- ¿Cuál es la desarrolladora que más críticas negativas tuvo y cuál es su juego más criticado?

Con los resultados que obtuvimos la empresa con mayores críticas negativas es “Ubisoft” con un total de 124 y usuarios negativos son 583, el juego de esta desarrolladora que obtuvo la mayoría de las críticas negativas fue “Figther within”.

- ¿Cuál es la desarrolladora que más críticas positivas tuvo más gustado?

  Según el análisis podemos saber que una extensión de Ubisoft logro tener las mejores críticas positivas esta es “Ubisoft Montreal” con un total de 1017 de críticas positivas y 3414 de usuarios positivos, el juego más gustado según la información es el “Far Cry 5”.

Hobbies y actividades

- ¿Cuál es un hobby que se puede practicar y que permite tener tiempo para estudiar y trabajar a medio tiempo?

  Basándonos en la información obtenida podemos decir que el deporte es la afición que les permite tener tiempo para otras cosas por ejemplo al estudiar los que practican deportes toman de 1 hora a 4 horas o más y el porcentaje de estos estudiantes que tiene trabajo a medio tiempo es de 50,98%.

- ¿Cómo influyen los hobbies que practicamos en la reducción del estrés y cómo pueden servir como indicadores de bienestar?

  Según vemos en el gráfico podemos notar que la afición que tiene más porcentaje de nivel de estrés en “bueno” es deportes con 16,67 % a si mismo esta posee un el mayor porcentaje en “fabuloso” con un 3,33 % lo que nos ayuda a concluir que si queremos tener muy bajo el estrés el deporte es la mejor manera de hacerlo.

Noticias mundiales

- ¿Cuáles son las tendencias mundiales en las noticias?

  El mayor tema del que se habla en las noticias es la “Política” con un total de 890 mil de este tema, siguiéndole el “Bienestar” con 699 mil y en tercer lugar los “Viajes” con 346 mil.

  En comparación con estas las menos tendencias es “Arte” con un de 1509.

Covid:

- ¿Cuál es la cantidad de casos confirmados, muertes y nuevos casos de COVID-19 en Brasil?

En Brasil se han registrado aproximadamente 2 millones de casos confirmados de COVID-19. Además, el número de muertes asciende a 88 mil, mientras que los casos activos son 508 mil.

NBA:

- ¿Cuál ha sido el mejor año de Lebrón James en términos de puntos y asistencias?

En 2003, durante su temporada de debut en la NBA, LeBron James logró un máximo de 43 puntos en un partido, destacándose por su capacidad anotadora desde el inicio. En total, acumuló 102 asistencias a lo largo de esa temporada. Estos datos reflejan su capacidad para anotar y asistir, aunque fueron solo un indicio del rendimiento más destacado que alcanzaría en sus años posteriores.

- ¿Cuál fue la suma total de puntos de los jugadores en la temporada 2021?

En la temporada 2021 de la NBA, la suma total de puntos anotados por todos los jugadores en conjunto fue de aproximadamente 50,000 puntos. Este dato refleja la magnitud del rendimiento ofensivo colectivo durante la temporada, destacando la contribución individual de cada jugador al total, así como el alto ritmo de juego característico de la liga en ese periodo.

STARBUCKS:

- ¿Cuál es el número total de locales de Starbucks con licencia en Chicago?

En Chicago, Starbucks cuenta con un total de 55 locales operados bajo el modelo de licencia. Este número refleja la presencia significativa de la cadena en la ciudad, destacando su expansión a través de asociaciones con licenciatarios locales que gestionan las tiendas bajo los estándares de la marca.

Liga española:

- ¿Cuáles son los puntos obtenidos por los equipos en Casa?
  En este caso el equipo con más puntos obtenidos en casa es Girona y el equipo con menos puntos realizados en casa es el Mallorca.

- ¿Cuáles son los puntos obtenidos por los equipos de visitante?

  En este caso el equipo con más puntos obtenidos de visitante es el Real Madrid y el equipo con menos puntos realizados de visitante es el Cadiz FC.

Netflix:

- ¿Cuáles son los tres directores con mayor cantidad de producciones realizadas?
  - Raul C.
  - Marcus R.
  - Jay K.
- ¿Cuáles son los tres países con más producciones realizadas?
  - United States
  - India
  - United Kingdom

Reseña Comida Peruana:

- ¿Cuáles son las 5 palabras más frecuentes?

  Las 5 palabras más frecuentes: comer, atender, recomendar, delicia y servicio.

11. Conclusiones y Recomendaciones

- **Conclusiones:**

  El proyecto de recopilación y análisis de datos ha demostrado la crucial importancia de la limpieza de datos para garantizar su calidad y precisión. Utilizando Jupyter Notebook, se facilitó un análisis interactivo y eficiente, mientras que la integración de múltiples bases de datos (MySQL Server, SQL Server, MongoDB) permitió aprovechar las ventajas específicas de cada una. La visualización a través de Power BI reveló patrones y tendencias significativas, mejorando la comunicación de hallazgos. Además, el proceso fomentó el aprendizaje continuo en habilidades técnicas y destacó la relevancia del análisis exploratorio para abrir nuevas preguntas e ideas en futuras investigaciones. En conjunto, estas etapas resaltan la necesidad de un enfoque integral en el manejo de datos para una toma de decisiones informada.

- **Recomendaciones:**

  Antes de crear una visualización de datos, es fundamental analizar y comprender la información contenida en los data sets. Este proceso implica identificar patrones, tendencias y posibles inconsistencias que puedan afectar la interpretación de los resultados. Al realizar un examen detallado de los datos,

se pueden detectar valores atípicos, datos faltantes o relaciones clave entre variables, lo que permite una representación gráfica más precisa y significativa. De esta manera, la visualización no solo será atractiva, sino que también facilitará una mejor toma de decisiones basada en información confiable y bien estructurada.

12. Desafíos y Problemas Encontrados

- Surgieron problemas con datos vacíos y errores de formato en especial en los .json, lo que requirió llenar esos datos nulos, para el formato se utilizó código en Python para solucionarlo.

- Al momento de general alguna visualización en PowerBi el programa no detectaba métricas en datos, se solucionó haciendo un recuento de cada una en una nueva columna.

- Al exportar el data sets de Covid no aceptaba la conexión para hacerlo, esto se debió a que el tipo de dato de la columna deaths/100 en algunos datos había incompatibilidad por lo que tuve que cambiar esos datos.

Referencias

[1] "Kaggle: your home for data science." https://www.kaggle.com/

[2] "Joker - movie reviews," *Rotten Tomatoes*. https://www.rottentomatoes.com/m/joker_2019/reviews

Anexos

Repositorio de Github:

Enlaces de los videos:

- Recopilación de datos: https://youtu.be/trll50zOnhs
- Conclusiones de los dashboards: https://youtu.be/D_miZNeMvzs

Enlaces a los dashboards en PowerBI:

Liga Española:
https://app.powerbi.com/view?r=eyJrIjoiMDczZTJhMjUtMmJjMi00MDlkLWJkYjYtZThjYzdmNmVlNDMyIiwidCI6IjY4MmE0ZTZhLWE3N2YtNDk1OC1hM2FjLTllMjY2ZDE4YWEzNyIsImMiOjR9

Netflix:
https://app.powerbi.com/view?r=eyJrIjoiZTA5YjJlZDctNmFlZC00MDk2LTkxNzItZTk4Yzg1M2JjNGIxIiwidCI6IjY4MmE0ZTZhLWE3N2YtNDk1OC1hM2FjLTllMjY2ZDE4YWEzNyIsImMiOjR9

Reseñas comida peruana:
https://app.powerbi.com/view?r=eyJrIjoiMGIwZTE5YjUtMWE3NC00ZjRlLWI2ZmMtNjhlYzljYTg2NDI2IiwidCI6IjY4MmE0ZTZhLWE3N2YtNDk1OC1hM2FjLTllMjY2ZDE4YWEzNyIsImMiOjR9

Videojuegos, Hobbies y Noticias:

https://app.powerbi.com/view?r=eyJrIjoiYzNiZjM2ODUtZjBmYi00OWQ5LTkzYWUtYmU3Yzk2YzJiMTUxIiwidCI6IjY4MmE0ZTZhLWE3N2YtNDk1OC1hM2FjLTllMjY2ZDE4YWEzNyIsImMiOjR9

Covid, NBA, Starbucks:

https://app.powerbi.com/view?r=eyJrIjoiZDM4MWQ2MDAtNmQ5Mi00MDNhLThkMjYtMjgzM2I4MGMxNzdlIiwidCI6IjY4MmE0ZTZhLWE3N2YtNDk1OC1hM2FjLTllMjY2ZDE4YWEzNyIsImMiOjR9