EQUIPO: LOS NUGGETS

Actividad en redes sociales: Lactancia materna y Fórmula

Valeria Margarita Espinoza Sánchez **18100168**

José Arcadio Rodríguez Matta **18100227**

INTRODUCCIÓN

Los medios masivos de comunicación son considerados como medios universales con el potencial de influir en las normas sociales.

Así, este estudio pretende explorar cuantitativa y cualitativamente la actividad que se presenta acerca de la alimentación mediante lactancia materna y mediante fórmula.

OBJETIVO

La realización del proyecto busca cumplir los siguientes objetivos:

- **Buscar el día con mayor actividad dentro de un plazo de 3 meses.**
- **Obtener el día de la semana con más publicaciones realizadas.**
- **Ver los países que generan mayor cantidad de contenido.**
- **Conocer la hora en la cual se genera mayor contenido relacionado.**

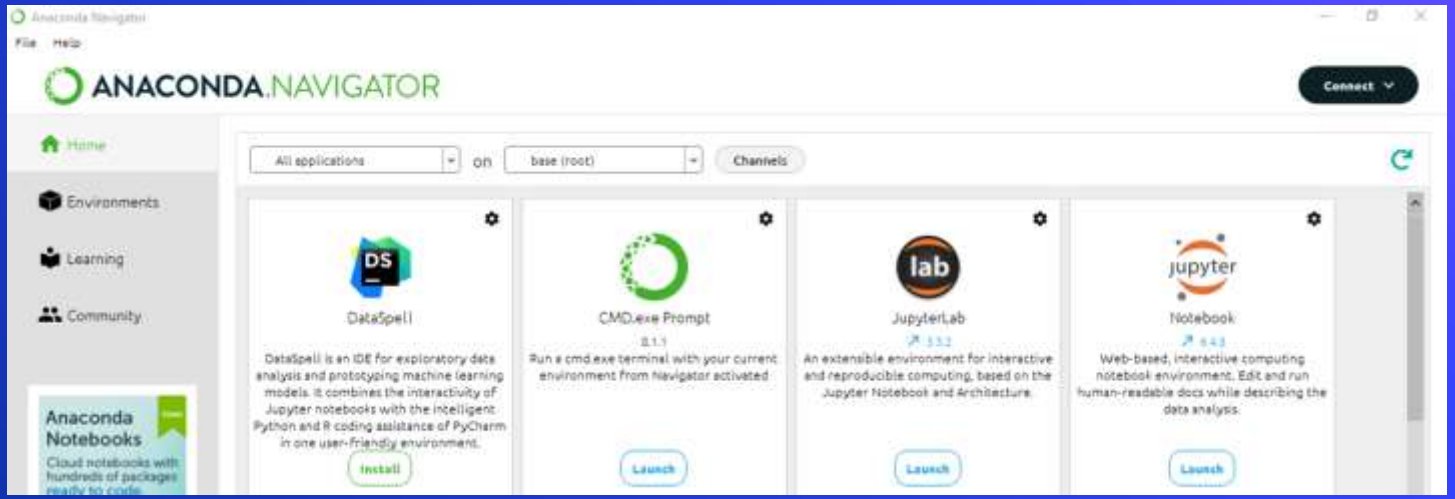
HERRAMIENTAS UTILIZADAS

- ➔ JUPYTERLAB
- ➔ PYTHON
- ➔ ANACONDA
- ➔ PANDAS
- ➔ NUMPY
- ➔ PLOTLY
- ➔ MATPLOTLIB

DESARROLLO

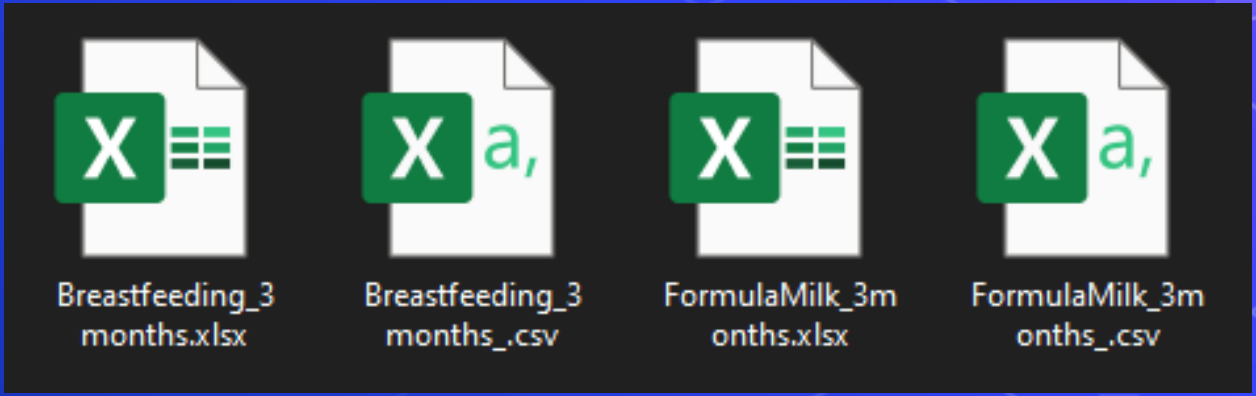
El proyecto de desarrolla en el siguiente entorno:

Anaconda, utilizando JupyterLab, escrito en Python con las librerías Pandas, Matplotlib, Numpy y Plotly para el muestreo y análisis de los datos



CONJUNTOS DE DATOS (DATASETS)

Los datasets obtenidos abarcan un periodo de tres meses, desde julio hasta septiembre del 2022



CONJUNTOS DE DATOS (DATASETS)

COLUMNAS

- ***status_id***: ID numérico único asignado a cada registro.
- ***created_at***: La fecha y hora en que se publicó
- ***text***: El texto del comentario/publicación
- ***display_text_width***: Largo del texto del post (número de caracteres)
- ***country***: País donde se publicó
- ***day***: Día de la semana en donde se publicó

A	B	C	D	E	F
status_id	created_at	text	display_text_width	country	day
1538998114270461952	2022-06-20 21:31:58 UTC	@TaylorLyonsMSJ HAHAHAHA. Sister, y	181	United States	lunes
1539010386506162176	2022-06-20 22:20:44 UTC	Gloria Dudney starting the #TNBFSym2	160	United States	lunes
1539026425377869824	2022-06-20 23:24:28 UTC	@KateNicholl @BelTel @SuzyJourno I'	74	United Kingdom	lunes
1539065335239819264	2022-06-21 01:59:05 UTC	on the fence, a part of me is DONE brea	117	United States	martes
1539100120696115200	2022-06-21 04:17:18 UTC	Lupig pag buntis aning breastfeeding u	55	Republic of the Philippines	martes
1539127189697970176	2022-06-21 06:04:52 UTC	@PoolsideGaGirl @Docmaker63 @Ka	76	United Kingdom	martes
1539128130929147904	2022-06-21 06:08:36 UTC	@PoolsideGaGirl @Docmaker63 @Ka	34	United Kingdom	martes
1539246405717573632	2022-06-21 13:58:35 UTC	....and suddenly I miss breastfeeding (	39	South Africa	martes
1539259459918106624	2022-06-21 14:50:27 UTC	Our breastfeeding journey has been ha	145	United States	martes

ANÁLISIS DE LOS DATASETS

Para el desarrollo del proyecto, primeramente se realizó la importación de las librerías numpy, pandas, plotly y malplotlib, que permitirán manipular datos y obtener información en base a ellos

[2]:
```
import numpy as numpy #librería de algebra lineal
import pandas as panda #librería para el procesamiento de datos
#librería para crear gráficas
import plotly.express as plotly #sofisticadas
import matplotlib.pyplot as matplot #sencillas
import re #librería para hacer uso de expresiones regulares

import os #se importan los datasets con los que se trabajaran
for dirname, _, filenames in os.walk('C:/Users/Valeria/Documents/9no Semestre/Analítica de Big Data/Proyecto Final/Breastfeeding/'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
C:/Users/Valeria/Documents/9no Semestre/Analítica de Big Data/Proyecto Final/Breastfeeding/Breastfeeding_3months.xlsx
C:/Users/Valeria/Documents/9no Semestre/Analítica de Big Data/Proyecto Final/Breastfeeding/Breastfeeding_3months_.csv
C:/Users/Valeria/Documents/9no Semestre/Analítica de Big Data/Proyecto Final/Breastfeeding/FormulaMilk_3months.xlsx
C:/Users/Valeria/Documents/9no Semestre/Analítica de Big Data/Proyecto Final/Breastfeeding/FormulaMilk_3months_.csv
```

ANÁLISIS DE LOS DATASETS

Al verificar la existencia de los archivos, se seleccionará el dataset con el cual se requiera trabajar

```
[4]: #Se realiza la lectura del archivo que se analizará
     dataset = panda.read_csv('C:/Users/Valeria/Documents/9no Semestre/Analítica de Big Data/Proyecto Final/Breastfeeding/Breastfeeding_3months_.csv

[5]: dataset.head() #Se muestra una parte de la información
```

```
[7]: #Se realiza la lectura del archivo que se analizará
     dataset = panda.read_csv('C:/Users/Valeria/Documents/9no Semestre/Analítica de Big Data/Proyecto Final/Breastfeeding/FormulaMilk_3months_.csv')

[8]: dataset.head() #Se muestra una parte de la información
```

ANÁLISIS DE LOS DATASETS

Posteriormente, como parte de la obtención de los datos, se comprueban los tipos de datos que contiene cada campo:

```
[6]: dataset.info() #Devuelve la información de las columnas

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 238045 entries, 0 to 238044
Data columns (total 6 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   status_id           238045 non-null  int64
 1   created_at          238045 non-null  object
 2   text                228058 non-null  object
 3   display_text_width  238045 non-null  int64
 4   country             485 non-null     object
 5   day                 238045 non-null  object
dtypes: int64(2), object(4)
memory usage: 10.9+ MB
```

```
[9]: dataset.info() #Devuelve la información de las columnas

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43559 entries, 0 to 43558
Data columns (total 6 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   status_id           41205 non-null  float64
 1   created_at          41205 non-null  object
 2   text                38195 non-null  object
 3   display_text_width  41205 non-null  float64
 4   country             71 non-null     object
 5   day                 41205 non-null  object
dtypes: float64(2), object(4)
memory usage: 2.0+ MB
```

ANÁLISIS DE LOS DATASETS

También se buscan la cantidad de registros que tengan alguna columna en blanco / nulos:

```
[7]: dataset.isna().sum() #Busca las columnas

[7]: status_id              0
     created_at             0
     text                9987
     display_text_width     0
     country           237560
     day                    0
     dtype: int64
```

```
[11]: dataset.isna().sum() #Busca las columnas

[11]: status_id              2354
      created_at             2354
      text                   5364
      display_text_width     2354
      country               43488
      day                    2354
      dtype: int64
```

ANÁLISIS DE LOS DATASETS

Para la columna **created_at**, se encontró que tiene un tipo de dato "**object**", por lo tanto hay que volver a comprobar con el método **type()**.

Se convierte a formato datetime:

```
[12]: #Devuelve el tipo de objeto en la columna 'created_at'
      type(dataset.at[0, 'created_at'])
      #El valor que devuelve es str

[12]: str

[13]: dataset['created_at'] = panda.to_datetime(dataset['created_at']) #Cambia de string datetime a datetime
```

ANÁLISIS DE LOS DATASETS

Ahora a manera informativa, se obtienen todos los países que se encuentren en el dataset (sin repetir, mediante el método unique() a la columna country):

```
[27]: dataset['country'].unique()

[27]: array([nan, 'United States', 'United Kingdom',
       'Republic of the Philippines', 'South Africa', 'Zimbabwe', 'India',
       'Rwanda', 'Canada', 'Ireland', 'Australia', 'Kenya', 'Vietnam',
       'Malaysia', 'East Timor', 'Denmark', 'Indonesia', 'Nigeria',
       'The Netherlands', 'Italy', 'Pakistan', 'Germany', 'Jamaica',
       'Switzerland', 'Norway', 'Singapore', 'Spain', 'Mexico',
       'Kingdom of Saudi Arabia', 'Ghana', 'Uganda', 'Thailand', 'Cyprus',
       'Trinidad and Tobago', 'Botswana', 'Portugal', 'Namibia',
       'New Zealand', 'Republic of Slovenia', 'Ethiopia',
       'United Arab Emirates', 'Latvia', 'Maldives', 'Zambia',
       'Dominican Republic', "People's Republic of China"], dtype=object)
```

```
[15]: dataset['country'].unique()

[15]: array([nan, 'United States', 'Australia', 'United Kingdom', 'Canada',
       'Republic of the Philippines', 'Botswana', 'Zimbabwe', 'Ghana',
       'Mexico', 'Pakistan', 'South Africa', 'Nigeria', 'Ireland'],
      dtype=object)
```

ANÁLISIS DE LOS DATASETS

Para el siguiente bloque se utiliza la librería PyCountry, una biblioteca de Python para completar los datos de los países para preparar el conjunto de datos para su comparación o agregación con otros.

```
import pycountry
mapeo = {country.name: country.alpha_3 for country in pycountry.countries} #alpha_3 cambia el nombre del país por sus 3 siglas
mapeo_banderas = {country.name: country.flag for country in pycountry.countries} #Obtiene las banderas de los países
```

ANÁLISIS DE LOS DATASETS

Al dataset se agregan dos columnas: **country_code** y **country_flag** y se llenan con el contenido de ***mapeo*** (códigos en tres siglas) y ***mapeo_banderas*** (imágenes de banderas) respectivamente.

```
[22]: dataset = dataset.assign(country_code = None) #Se agrega la columna country_code
      dataset = dataset.assign(country_flag = None) #Se agrega la columna country_flag
      for index in dataset.index:   #Realiza recorrido por filas del dataset
          country = dataset.at[index, 'country']  #country es igual al elemento en index de la columna country
          dataset.at[index, 'country_code'] = mapeo.get(country) #Se coloca en el dataset las siglas del país
          dataset.at[index, 'country_flag'] = mapeo_banderas.get(country) #Se coloca en el dataset las banderas
```

ANÁLISIS DE LOS DATASETS

Se comprueba dicha información consultando las dos columnas con unique()

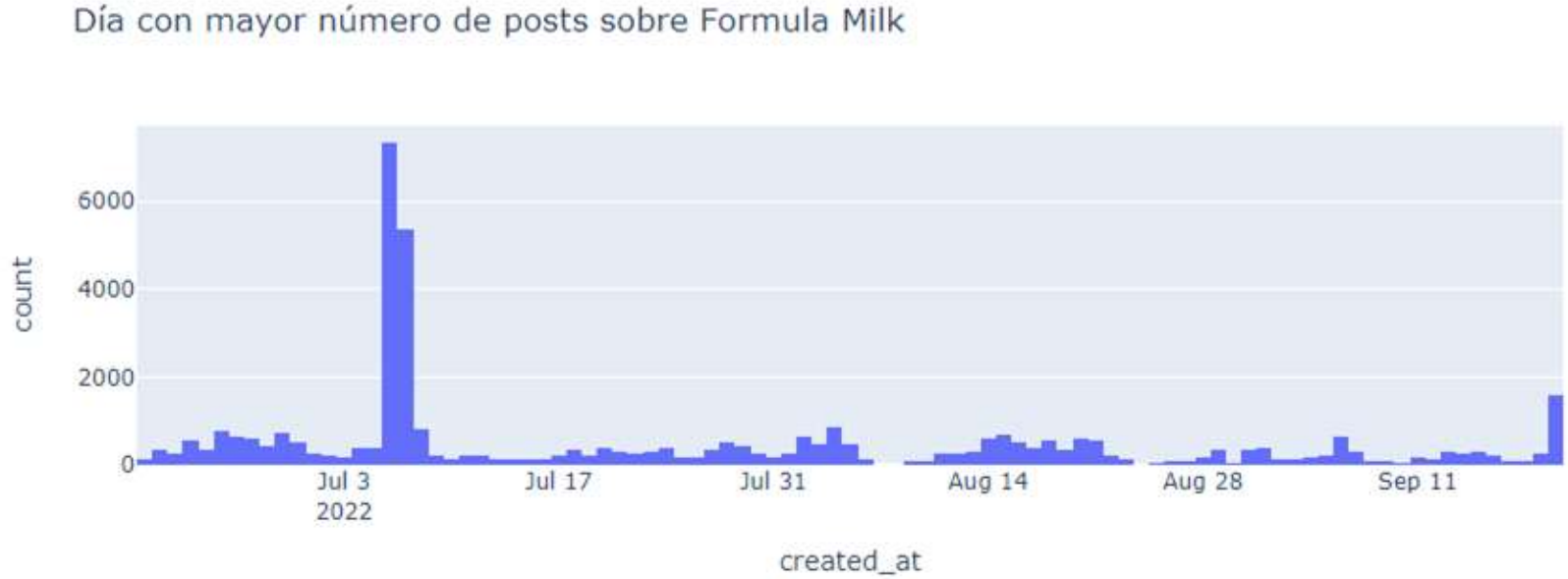
```
[45]: dataset['country_code'].unique() #Obtiene de la columna country_code los elementos no repetidos, siendo 45 países

[45]: array([None, 'USA', 'GBR', 'PHL', 'ZAF', 'ZWE', 'IND', 'RWA', 'CAN',
       'IRL', 'AUS', 'KEN', 'VNM', 'MYS', 'TLS', 'DNK', 'IDN', 'NGA',
       'NLD', 'ITA', 'PAK', 'DEU', 'JAM', 'CHE', 'NOR', 'SGP', 'ESP',
       'MEX', 'SAU', 'GHA', 'UGA', 'THA', 'CYP', 'TTO', 'BWA', 'PRT',
       'NAM', 'NZL', 'SVN', 'ETH', 'ARE', 'LVA', 'MDV', 'ZMB', 'DOM',
       'CHN'], dtype=object)

[46]: dataset['country_flag'].unique() #Obtiene de la columna country_code los elementos no repetidos, siendo 45 países

[46]: array([None, 'US', 'GB', 'PH', 'ZA', 'ZW', 'IN', 'RW', 'CA', 'IE', 'AU',
       'KE', 'VN', 'MY', 'TL', 'DK', 'ID', 'NG', 'NL', 'IT', 'PK', 'DE',
       'JM', 'CH', 'NO', 'SG', 'ES', 'MX', 'SA', 'GH', 'UG', 'TH', 'CY',
       'TT', 'BW', 'PT', 'NA', 'NZ', 'SI', 'ET', 'AE', 'LV', 'MV', 'ZM',
       'DO', 'CN'], dtype=object)
```

ANÁLISIS DE LOS DATASETS

Con el siguiente bloque se crea un histograma, que permite ver el día con mayor número de posts generados. Para ello se utiliza Plotly con el método .histogram

```
figura = plotly.histogram(dataset, x = 'created_at', title='Día con mayor número de posts sobre Breastfeeding')
figura.show()
```

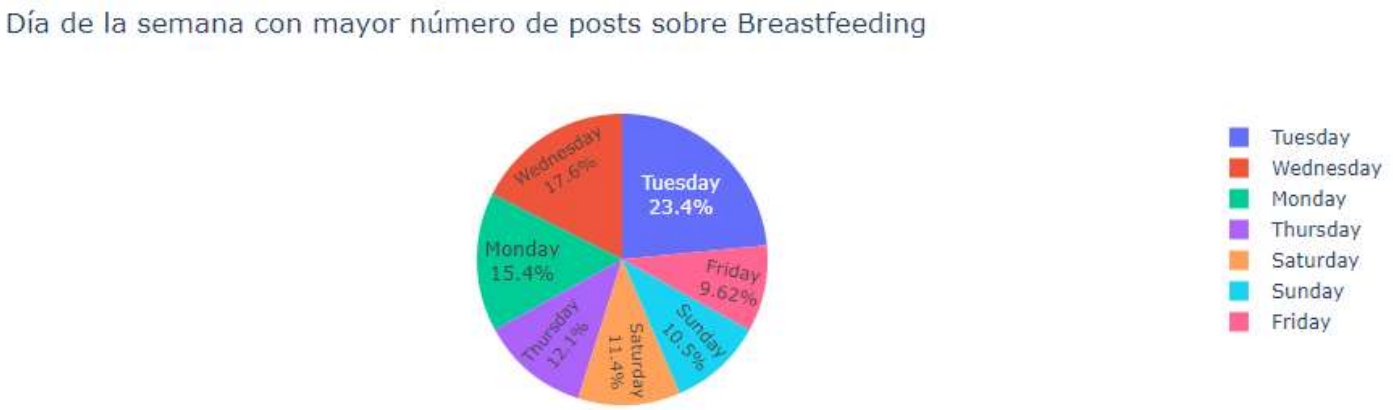
ANÁLISIS DE LOS DATASETS



ANÁLISIS DE LOS DATASETS

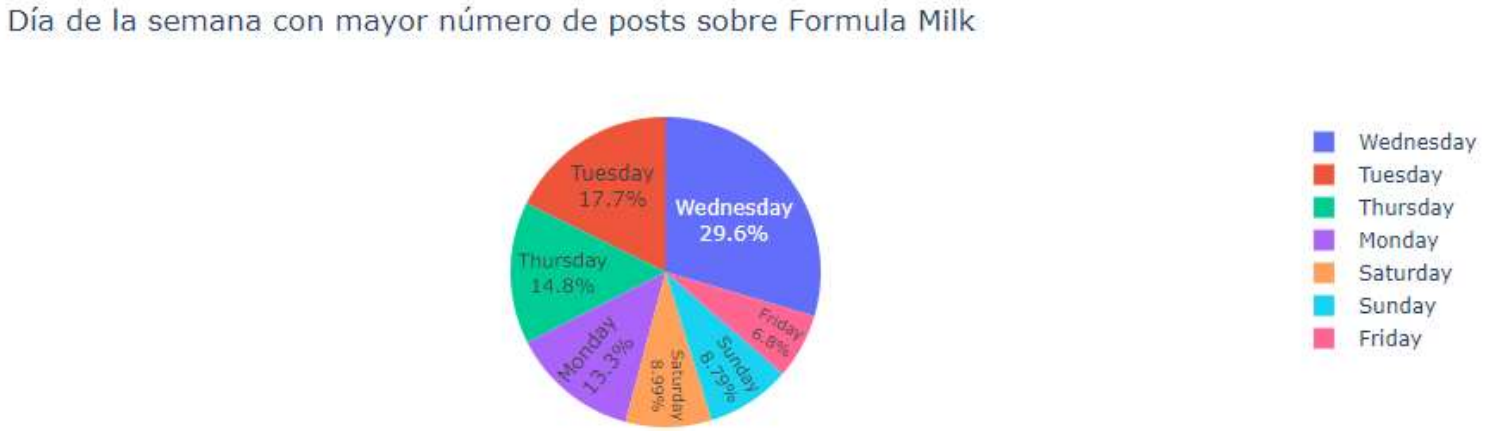
Con la siguiente gráfica se hace un análisis sobre el día de la semana con mayor número de posts, para ello:

```
[92]: pie = dataset.groupby(['day']).size().reset_index(name = 'size')
      pie.sort_values(by='size', ascending=False ,inplace=True)
      figura = plotly.pie(pie, values='size', names='day',
                  title='Día de la semana con mayor número de posts sobre Breastfeeding',
                  hover_data=['size'])
      figura.update_traces(textposition='inside', textinfo='percent+label')
      figura.show()
```



ANÁLISIS DE LOS DATASETS

```
[41]: pie = dataset.groupby(['day']).size().reset_index(name = 'size')
      pie.sort_values(by='size', ascending=False ,inplace=True)
      figura = plotly.pie(pie, values='size', names='day',
                  title='Día de la semana con mayor número de posts sobre Formula Milk',
                  hover_data=['size'])
      figura.update_traces(textposition='inside', textinfo='percent+label')
      figura.show()
```



ANÁLISIS DE LOS DATASETS

En el siguiente bloque de código se genera un conjunto de datos de los países con mayor número de publicaciones realizadas. Para ello:

[85]:
```
dtMundial = dataset.groupby(['country', 'country_code', 'country_flag']).size().reset_index(name = 'size')
dtMundial.sort_values(by='size', ascending=False ,inplace=True)
```

[86]:
```
dtMundial.head()
```

[86]:

	country	country_code	country_flag	size
41	United States	USA	US	213
40	United Kingdom	GBR	GB	81
15	Kenya	KEN	KE	23
32	South Africa	ZAF	ZA	18
10	India	IND	IN	17

ANÁLISIS DE LOS DATASETS

[34]:
```
dtMundial = dataset.groupby(['country', 'country_code', 'country_flag']).size().reset_index(name = 'size')
dtMundial.sort_values(by='size', ascending=False ,inplace=True)
```

[35]:
```
dtMundial.head()
```

[35]:

	country	country_code	country_flag	size
11	United States	USA	US	42
10	United Kingdom	GBR	GB	10
3	Ghana	GHA	GH	5
8	Philippines	PHL	PH	4
0	Australia	AUS	AU	2

ANÁLISIS DE LOS DATASETS

Al generar la agrupación, se crea la gráfica de un **mapa mundial** que mostrará la ubicación geográfica de los países donde se hacen más publicaciones, utilizando **size** para poder graficar sobre el mapa.

```
[87]: figura = plotly.scatter_geo(dtMundial, locations="country_code", color="country",
                                  hover_name="country", size="size",
                                  projection="natural earth1",title="Países con más publicaciones de Breastfeeding")
      figura.show()
```

ANÁLISIS DE LOS DATASETS

[37]:
```
figura = plotly.scatter_geo(dtMundial, locations="country_code", color="country",
                    hover_name="country", size="size",
                    projection="natural earth1", title="Países con más publicaciones de Formula Milk")
figura.show()
```

ANÁLISIS DE LOS DATASETS

Para el siguiente bloque de código, se crea una agrupación de la cantidad de publicaciones que se realizan en cada hora de un día, tomando en cuenta:

```
[88]: dt=dataset.groupby([dataset['created_at'].dt.hour])['status_id'].count()
      print(dt)
```

```
created_at
0      7102
1      7259
2      7226
3      7000
4      7095
5      7431
6      8300
7      9714
8      9202
9      9171
10     9375
11    11064
12    12398
13    13030
14    13734
15    14172
16    13775
17    12183
18    11280
19    10814
20    10256
21     9263
22     8817
23     8384
Name: status_id, dtype: int64
```

ANÁLISIS DE LOS DATASETS

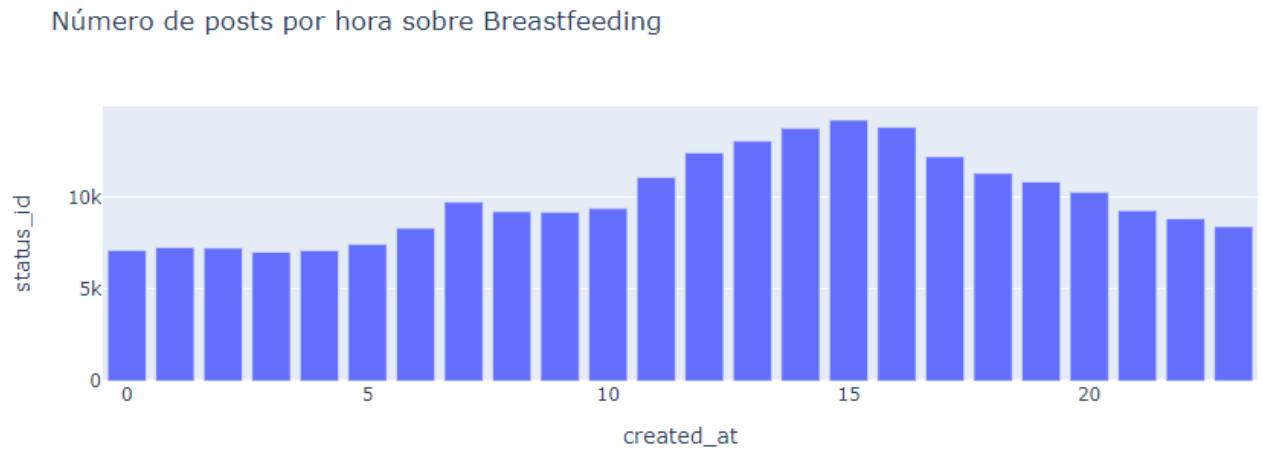
```
[25]: dt=dataset.groupby([dataset['created_at'].dt.hour])['status_id'].count()
      print(dt)
```

```
created_at
0.0     1437
1.0     1452
2.0     1407
3.0     1345
4.0     1322
5.0     1158
6.0     1284
7.0     1330
8.0     1235
9.0     1332
10.0    1272
11.0    1594
12.0    1697
13.0    1833
14.0    2001
15.0    2161
16.0    2135
17.0    2009
18.0    1825
19.0    1925
20.0    3288
21.0    2404
22.0    2054
23.0    1705
Name: status_id, dtype: int64
```

ANÁLISIS DE LOS DATASETS

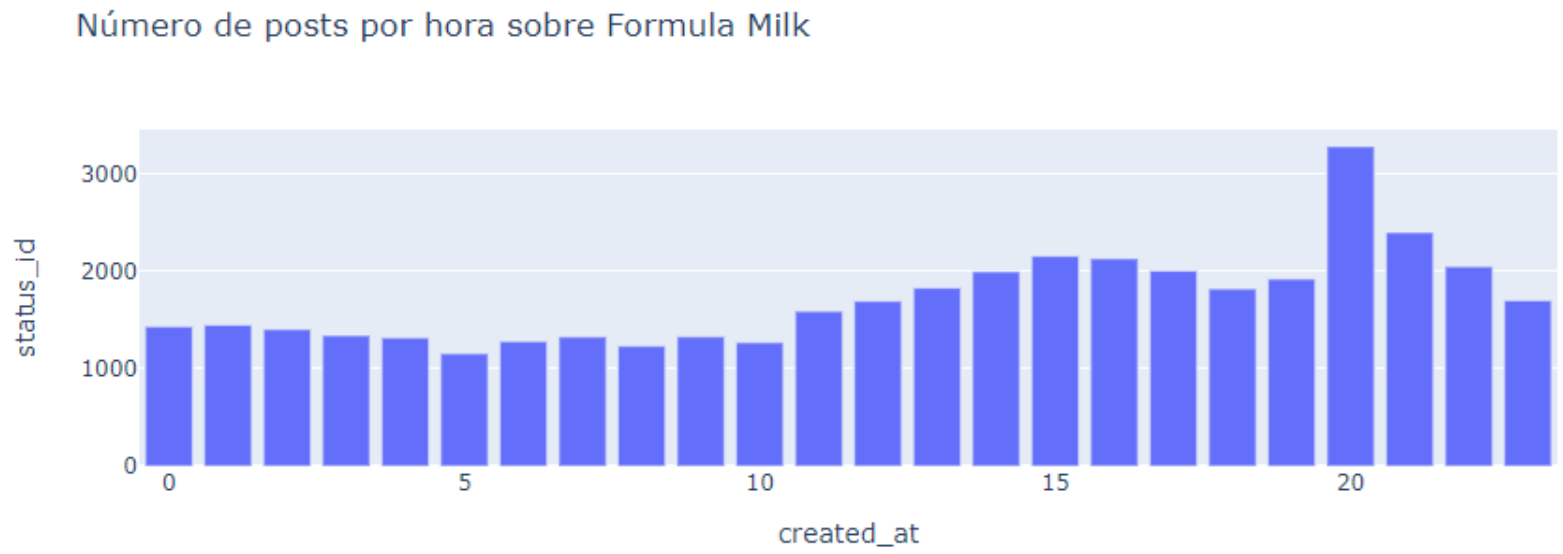
Se genera la gráfica de barras que permite ver las horas del día con mayor número de publicaciones, utilizando status_id como los números a graficar y created_at como las 24 horas del día.

```
[90]: figura = plotly.bar(dt, y='status_id', title="Número de posts por hora sobre Breastfeeding")
      figura.show()
```



ANÁLISIS DE LOS DATASETS

```
[39]: figura = plotly.bar(dt, y='status_id', title="Número de posts por hora sobre Formula Milk")
      figura.show()
```



COMPARATIVA DE RESULTADOS

Para tener una mejor visión de los resultados obtenidos, se realizó una tabla comparativa de ambos temas con respecto a las gráficas realizadas.

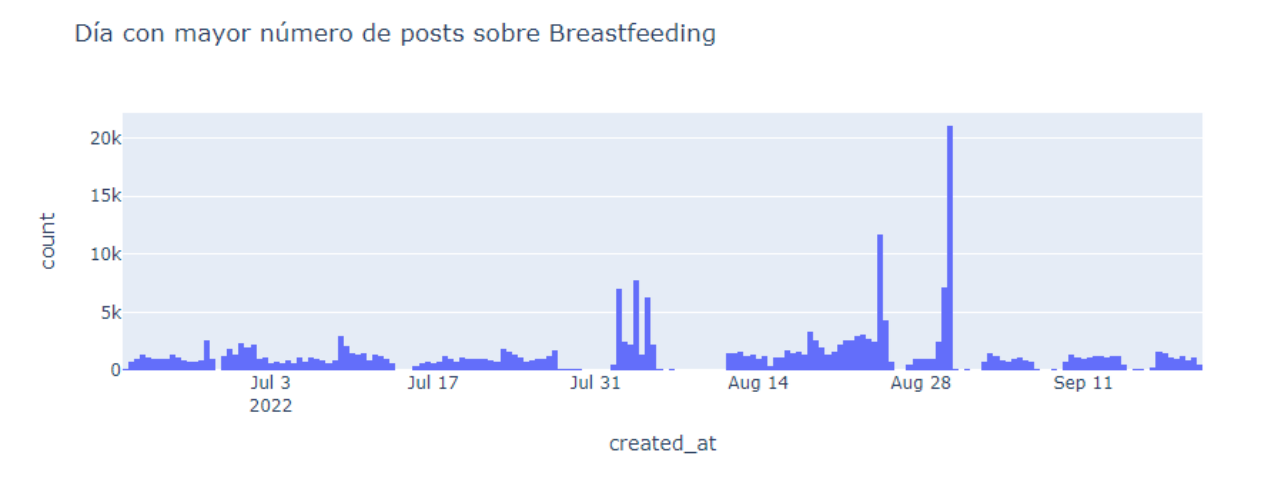
Gráfica	Breastfeeding	Formula Milk
Fecha con mayor número de posts	30-Agosto con 21,062 publicaciones	05-Julio con 7,343 publicaciones
Día de la semana con mayor número de posts	Martes con 23.4% con 55,812 publicaciones	Miércoles con 29.6% con 12,195 publicaciones
Países con mayor número de Posts	Estados Unidos us	Estados Unidos us
Número de posts por hora	3:00 pm con 14,172	8:00 pm horas con 3,288

ANÁLISIS DE RESULTADOS OBTENIDOS POR GRÁFICA

Fecha con mayor número de posts realizados

El histograma realizado representa todos los días que se realizaron publicaciones dentro de un periodo de tres meses, en este caso, los meses abarcan desde el 20 de junio hasta el 20 de septiembre, donde el eje x es la fecha de creación y el eje y la cantidad de posts realizados.

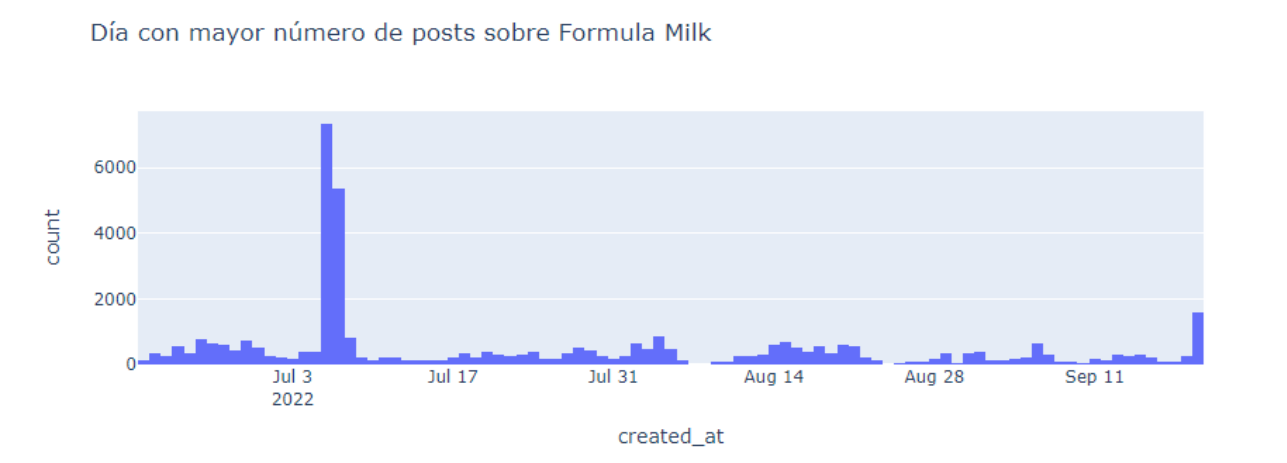
- Para Breastfeeding, el día con mayor cantidad de publicaciones realizadas es el día 30 de agosto.



ANÁLISIS DE RESULTADOS OBTENIDOS POR GRÁFICA

Fecha con mayor número de posts realizados

- **Para Formula Milk, el día con mayor cantidad de publicaciones realizadas es el día 05 de julio.**



ANÁLISIS DE RESULTADOS OBTENIDOS POR GRÁFICA

Fecha con mayor número de posts realizados

Al realizar el análisis de ambos histogramas, se llegó a las siguientes conclusiones:

- El tema de Breastfeding tuvo mayor actividad debido a que la primera semana de agosto es la semana para la lactancia materna. En general, mantuvo un nivel consistente de publicaciones.
- El tema de Formula Milk no tuvo tanta actividad, a excepción de un par de días de julio.

ANÁLISIS DE RESULTADOS OBTENIDOS POR GRÁFICA

Día de la semana con mayor número de posts realizados

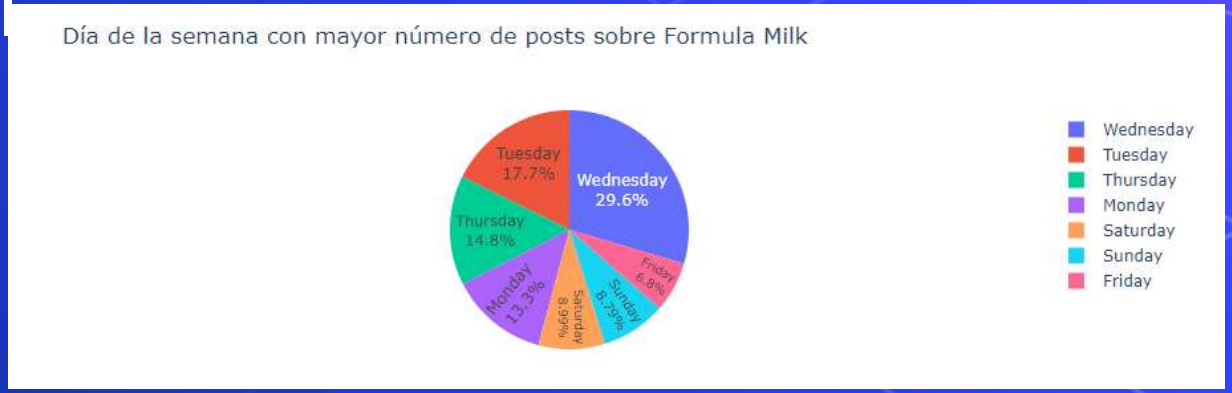
La gráfica de pie está formada por 7 partes, donde cada una de ellas representa cada uno de los días de la semana, con su porcentaje de la cantidad de posts realizados.

- **Para Breastfeeding, el día de la semana con mayor cantidad de publicaciones realizadas es el día Martes.**

ANÁLISIS DE RESULTADOS OBTENIDOS POR GRÁFICA

Día de la semana con mayor número de posts realizados

- Para Formula Milk, el día de la semana con mayor cantidad de publicaciones realizadas es el día Miércol



ANÁLISIS DE RESULTADOS OBTENIDOS POR GRÁFICA

Día de la semana con mayor número de posts realizados

Al realizar el análisis de ambas gráficas de pie, se llegó a las siguientes conclusiones:

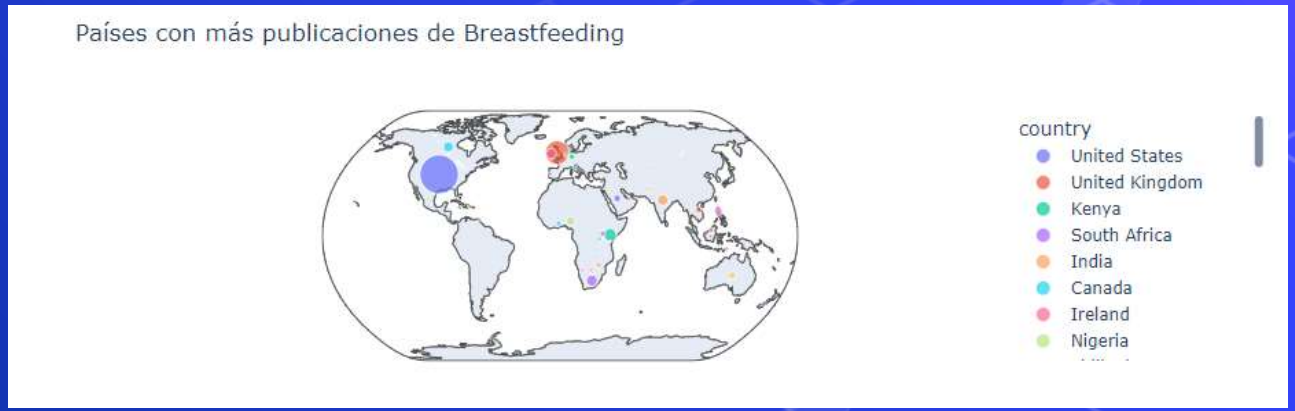
- En ambos tópicos, los días con menos actividad son los viernes, sábados y domingos.
- Ambos días de la semana que tienen más actividad, representan un poco más de ⅕ de la gráfica.

ANÁLISIS DE RESULTADOS OBTENIDOS POR GRÁFICA

Países con mayor cantidad cantidad de publicaciones realizadas

La gráfica muestra un mapa mundial que ubica los países que realizaron la mayor cantidad de posts. Además, la cantidad de posts está representada por un círculo, mientras más grande sea, es mayor el número de publicaciones realizadas.

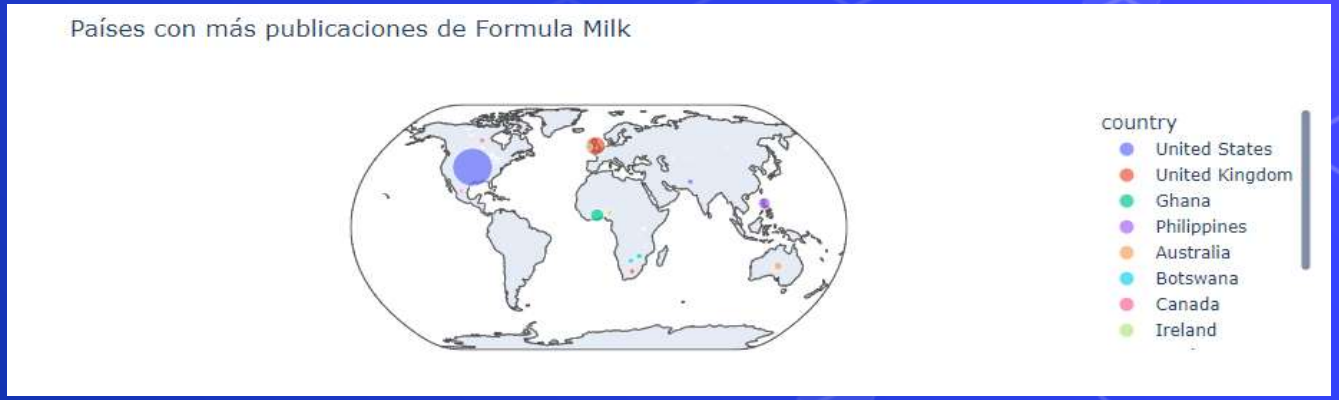
- **Para Breastfeeding, el país con mayor cantidad de publicaciones realizadas es Estados Unidos.**



ANÁLISIS DE RESULTADOS OBTENIDOS POR GRÁFICA

Países con mayor cantidad de publicaciones realizadas

- Para Formula Milk, el país con mayor cantidad de publicaciones realizadas es Estados Unidos.



ANÁLISIS DE RESULTADOS OBTENIDOS POR GRÁFICA

Países con mayor cantidad cantidad de publicaciones realizadas

Al realizar el análisis de ambas gráficas, se llegó a las siguientes conclusiones:

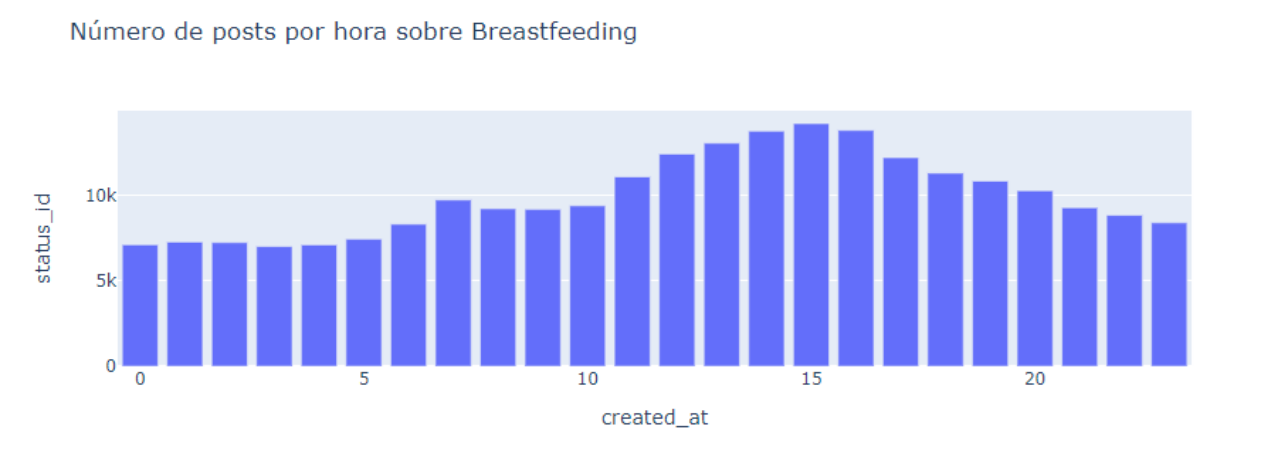
- **Los primeros dos países con mayor actividad para ambos temas son Estados Unidos y Reino Unido**

ANÁLISIS DE RESULTADOS OBTENIDOS POR GRÁFICA

Horas con mayor cantidad de publicaciones realizadas

La gráfica de barras representa está conformada por las 24 horas de un día, donde el eje x son las horas y el eje y es la cantidad de publicaciones que se realizaron.

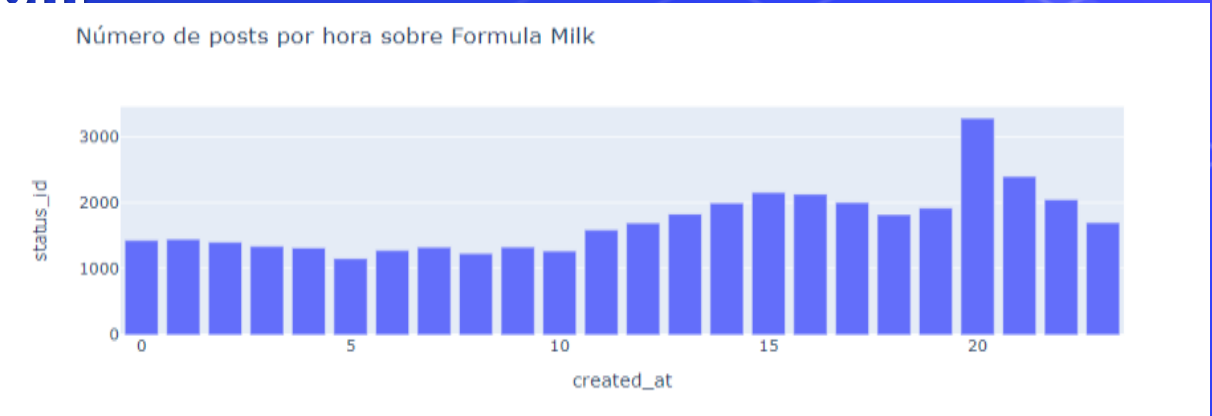
- Para Breastfeeding, la hora con mayor cantidad de publicaciones realizadas es las 03:00 pm.



ANÁLISIS DE RESULTADOS OBTENIDOS POR GRÁFICA

Horas con mayor cantidad de publicaciones realizadas

- Para Formula Milk, la hora con mayor cantidad de publicaciones realizadas es las 08:00 pm



ANÁLISIS DE RESULTADOS OBTENIDOS POR GRÁFICA

Horas con mayor cantidad de publicaciones realizadas

Al realizar el análisis de ambas gráficas, se llegó a las siguientes conclusiones:

- Para el tema de Breastfeeding, las horas con mayor actividad son entre las 02:00 pm y 04:00 pm.
- Para el tema de Formula Milk, solamente se tiene una hora con mayor actividad, siendo las 08:00 pm.
- Breastfeeding mantiene un nivel de actividad un poco más constante que Formula Milk.

OBSERVACIONES FINALES

Como observaciones finales, el tema de Breastfeeding tiene mayor actividad dentro de las redes sociales a comparación de Formula Milk ya que es un tema del cual no se suele hablar mucho con las personas cercanas, provocando que inicien una conversación sobre el mismo en la cual generan preguntas y comparten experiencias en sus procesos.

En general, ambos métodos son efectivos para la alimentación y nutrición de los bebés, pero sin duda alguna, debido al aumento de popularidad del tema de Breastfeeding es que se genera mayor actividad de la misma.