



**Tecnológico
de Monterrey**

**TC1002s. 600 Herramientas computacionales: el arte de la
analítica.**

Actividad Evaluable: Obtención de estadísticas descriptivas

Profesor : Fabiola Díaz.

Estudiantes :

Jatzive Adriana Pérez Solís | A01701715

Cutberto Arizabalo Nava | A01411431

Diego Carrillo Torres | A01612532

La base de datos que nosotros elegimos se llama “Trending Youtube Video Statistics”, la cual nos muestra los videos más populares en la plataforma, está dividida por canales, likes, dislikes, comentarios, los temas que dominan en los videos, etc.

Esta base de datos está dividida por países y nosotros elegimos los datos que corresponden a México.

link: <https://www.kaggle.com/datasnaek/youtube-new?select=USvideos.csv>

1. Cargar los datos usando tu lector de csv o con pandas. Es recomendable hacerlo con pandas.

Punto 1

```
In [3]: import pandas as pd
data=pd.read_csv('MXvideosUTF8.csv')
```

```
In [4]: data.head()
```

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes
0	SbOwzAI9ZfQ	17.14.11	Capítulo 12 MasterChef 2017	MasterChef 2017	24	2017-11-13T06:06:22.000Z	MasterChef Junior 2017 TV Azteca recetas ...	310130	4182	3
1	kiOV6Xh-DnI	17.14.11	ALEXA EX-INTEGRANTE DEL GRUPO TIMBIRICHE RENUN...	Micky Contreras Martinez	22	2017-11-13T05:11:58.000Z	La Voz Mexico 7	104972	271	1
2	6L2ZF7Qzsbk	17.14.11	LOUIS CKAGÁ - EL PULSO DE LA REPÚBLICA	El Pulso De La República	25	2017-11-13T17:00:02.000Z	Chumel Torres El Pulso de la Republica noti...	136064	10105	2
3	hcY52MFWMMDM	17.14.11	Sismo de 6.7 sacude Costa Rica 12 Noviembre 2017	Casanare	25	2017-11-13T03:47:10.000Z	temblor costa rica sismo en costa rica	96153	378	1
4	_OXDcGPVaa4	17.14.11	DOG HACKS MUSAS LESSIE LOS POLINESIOS	Musas	26	2017-11-13T19:17:48.000Z	MUSAS lessie karen hacks perros dogs...	499965	57781	6

2. Verificar la cantidad de datos que tienen, las variables que contiene cada vector de datos e identificar el tipo de variables.

```
name: video_id, length: 40451, dtype: o
```

```
In [24]: i = 0
for columns in data:
    print(data.columns[i])
    i+=1
```

```
video_id
trending_date
title
channel_title
category_id
publish_time
tags
views
likes
dislikes
comment_count
thumbnail_link
comments_disabled
ratings_disabled
video_error_or_removed
description
```

Para conocer las columnas de datos

```
In [25]: data.info
```

```
Out[25]: <bound method DataFrame.info of          video_id trending_date \
0      Sb0wzA19ZfQ      17.14.11
1      k1OV6Xh-DnI      17.14.11
2      6L2ZF7Qzsbk      17.14.11
3      hcY52MFWMMDM      17.14.11
4      _OXDcGPVAa4      17.14.11
...      ...
40446 r63VB0agGAo      18.14.06
40447 i7r_kMbyngk      18.14.06
40448 _jnwjdMe3Zo      18.14.06
40449 pAH9omNAWA4      18.14.06
40450 dj5Z4jTE3-c      18.14.06

          title \
0      Capítulo 12 | MasterChef 2017
1      ALEXA EX-INTEGRANTE DEL GRUPO TIMBIRICHE RENUN...
2      LOUIS CKAGÃ“ - EL PULSO DE LA REPÃ“BLICA
3      Sismo de 6.7 sacude Costa Rica 12 Noviembre 2017
4      DOG HACKS | MUSAS LESSLIE LOS POLINESIOS
...      ...
40446 Shawn Mendes x Portugal (FPF Official World Cu...
40447 AMLO llega con su esposa al Tercer Debate en ...
40448 Maire usa una blusa kawaii Ã“ adorable o ridÃ“...
40449 La Jefa del CampeÃ“n - CapÃ“tulo 2 Parte 3/4
40450 Ã“POR QUÃ“ SHANKS ES TAN RESPETADO POR TODOS E...

          channel_title  category_id  publish_time \
0      MasterChef 2017          24  2017-11-13T06:06:22.000Z
1      Micky Contreras Martinez          22  2017-11-13T05:11:58.000Z
2      El Pulso De La RepÃ“blica          25  2017-11-13T17:00:02.000Z
3      Casanare          25  2017-11-13T03:47:10.000Z
4      Musas          26  2017-11-13T19:17:48.000Z
```

Para ver las variables y cantidad de datos de cada vector.

Hay 40450 registros por cada vector, 16 vectores de datos

Tipos de dato:

- Category_id, views, likes, dislikes, comment_count = int
- trending_date, publish_time = date
- title, channel_title, tags, thumbnail_link, description = string
- ratings_disabled, video_error_or_removed = bool

3. Analizar las variables para saber qué representa cada una y en qué rangos se encuentran. Si la descripción del problema no lo indica, utiliza el máximo y el mínimo para encontrarlo.

Las variables que están presentes en el análisis son: "Category_id" que representa el id del video, "views" que son las veces que el video ha sido reproducido, "likes y dislikes" que por medio de este rubro se identifica el apoyo o interés que tuvieron los usuarios con el video, "comment_count" representa las veces que ha sido comentado el video.

De igual manera podemos ver en la gráfica, los rangos en los que se encuentran por medio del máximo y mínimo.

```
In [28]: data.describe()
```

```
Out[28]:
```

	category_id	views	likes	dislikes	comment_count
count	40451.000000	4.045100e+04	4.045100e+04	4.045100e+04	40451.000000
mean	21.003140	3.423820e+05	1.586184e+04	7.471604e+02	2039.660008
std	5.878995	1.714691e+06	8.108987e+04	1.095358e+04	13938.031797
min	1.000000	1.570000e+02	0.000000e+00	0.000000e+00	0.000000
25%	20.000000	1.681300e+04	2.990000e+02	1.700000e+01	42.000000
50%	24.000000	5.697300e+04	1.246000e+03	6.300000e+01	196.000000
75%	24.000000	2.068940e+05	7.226000e+03	2.670000e+02	885.000000
max	43.000000	1.009124e+08	4.470923e+06	1.353667e+06	905925.000000

4. Basándose en la media, mediana y desviación estándar de cada variable, qué conclusiones puedes entregar de los datos.

Podemos ver de acuerdo a la desviación estándar que la cantidad de vistas, likes, dislikes y comentarios entre los videos de la base de datos es muy alta, lo que indica que los videos analizados tuvieron un impacto muy distinto al compararlos unos con otros. Esto quiere decir que tenemos una base de datos bastante heterogénea y que no únicamente cuenta con registros de videos virales o de videos que no tuvieron impacto. Por lo tanto, el promedio de vistas, likes y comentarios nos puede ser útil para conocer el impacto promedio que un video de youtube puede tener, ya que se considera tanto los videos virales como videos que solo llegaron a 100 visualizaciones y cero interacciones (likes, dislikes y comentarios).

Repositorio:

[Fdiaz17/TC1002600 at SemanaTec8 \(github.com\)](https://github.com/Fdiaz17/TC1002600)