Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Querétaro

Herramientas computacionales: el arte de la analítica (Gpo 600)

Equipo 3

Actividad Evaluable: CSV

Reto analítica

Profesora

Fabiola Díaz Nieto

Presenta

Eliuth Balderas Neri - A01703315

Yamil Martínez López - A01707385

Diego Iturbe Bravo - A01708272

17 de marzo de 2021

La base de datos que se utilizó en nuestro trabajo fue Kaggle, la cúal nos mostraba varias bases de datos como por ejemplo plataformas de video que en nuestro caso fue una que escogimos, esta nos muestra algunas series de Prime Video, donde se nos muestra el Número de serie, Nombre del programa, el año de lanzamiento, número de temporadas,género, rating, edad promedio de los espectadores, idioma. Esta base de datos corresponde a este link: https://www.kaggle.com/nilimajauhari/amazon-prime-tv-shows

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

   Se utilizó Pandas para el lector de los datos csv. En esta ocasión escogimos una base de datos que registra las series en streaming actualmente de Amazon Prime Video.

```
In [2]: import pandas
        df = pandas.read_csv('Prime Tv Shows.csv')
        print(df)
```

```
     S.no.                          Name of the show  Year of release  \
0        1                                Pataal Lok           2020.0
1        2                                    Upload           2020.0
2        3                The Marvelous Mrs. Maisel            2017.0
3        4                   Four More Shots Please            2019.0
4        5                                  Fleabag            2016.0
..     ...                                       ...              ...
390    391  The 2018 Rose Parade Hosted by Cord & Tish           2018.0
391    392                Aliens Love Underpants And…           2017.0
392    393          Gina Brillon: The Floor is Lava             2020.0
393    394                                       NaN              NaN
394    395                                       NaN              NaN

     No of seasons available Language          Genre  IMDb rating  \
0                        1.0    Hindi          Drama          7.5
1                        1.0  English  Sci-fi comedy          8.1
2                        3.0  English  Drama, Comedy          8.7
3                        2.0    Hindi  Drama, Comedy          5.3
4                        2.0  English         Comedy          8.7
..                       ...      ...            ...          ...
390                      1.0  English         Comedy          NaN
391                      1.0  English           Kids          NaN
392                      1.0  English         Comedy          NaN
393                      NaN      NaN            NaN          NaN
394                      NaN      NaN            NaN          NaN

    Age of viewers
0              18+
1              16+
2              16+
3              18+
4              18+
..             ...
390            All
391            All
392            16+
393            NaN
394            NaN

[395 rows x 8 columns]
```

2. **Verifica la cantidad de datos que tienen, las variables que contiene cada vector de datos e identifica el tipo de variables.**

El código nos marca que tenemos una cantidad de ocho(8) variables, las cuales vendrían a ser cada columna. Al igual que 393 renglones, los cuales nos permiten calcular un total de 3,144 datos totales.

```
[393 rows x 8 columns]
```

Las variables de nuestra base de datos son: Número de serie, Nombre del programa, el año de lanzamiento, número de temporadas,género, rating, edad promedio de los espectadores, idioma. Estas variables pudieron ser obtenidas a través de Python:

```
In [9]: i=0
        for colums in df:
            print (base.columns[i])
            i=i+1

S.no.
Name of the show
Year of release
No of seasons available
Language
Genre
IMDb rating
Age of viewers
```

Nuestras variables fueron identificadas como tipo de dato:

INT - Número de serie, número de temporadas, rating.

DATE - Año de lanzamiento.

STRING - Nombre del programa, género, idioma, (edad promedio de los espectadores) este dato supimos que era tipo string porque no nos

regresó máximos ya que al tener un de simbología el signo de '+' alteró el tipo de dato.

```
In [25]: base['Age of viewers'].describe()

Out[25]: count      393
         unique       5
         top        16+
         freq       150
         Name: Age of viewers, dtype: object
```

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

   A partir de un análisis de las variables pudimos determinar el significado de cada variable individualmente.

   Nuestra primera variable "S.no." nos indica el número de serie en orden de lanzamiento en la plataforma. La variable "Name of the show" nos da el nombre de la serie.

   La variable "Year of release" nos indica la fecha en la que se estrenó la serie, independientemente a Amazon Prime, y tiene una rango desde el año 1926 a el 2020.

   La variable "No of seasons available" nos da la cantidad de temporadas que tiene cada serie disponible en la plataforma y tiene un rango de 1 a 20 temporadas.

   Nuestra variable "Language" indica el lenguaje original con el que se grabó la serie.

La variable “Genre” indica el género al que pertenece cada serie.

La variable “IMDb rating” nos indica la calificación de la serie en la plataforma de cinematografía IMDb y tiene un rango de 3.7 a 9.5 de calificación.

Por último, la variable “Age of viewers” nos indica la edad promedio de los espectadores de la serie en Amazon Prime, sin embargo, no se pudo obtener un rango ya que los datos se encuentran en tipo string.

Se utilizó el siguiente código para obtener los minimos y maximos de nuestras variables INT :

```
In [22]: base['Year of release'].describe()

Out[22]: count     393.000000
         mean     2011.274809
         std        12.943787
         min      1926.000000
         25%      2011.000000
         50%      2016.000000
         75%      2018.000000
         max      2020.000000
         Name: Year of release, dtype: float64
```

```
In [23]: base['No of seasons available'].describe()

Out[23]: count    393.000000
         mean       2.608142
         std        2.592008
         min        1.000000
         25%        1.000000
         50%        2.000000
         75%        3.000000
         max       20.000000
         Name: No of seasons available, dtype: float64
```

```
In [24]: base['IMDb rating'].describe()

Out[24]: count    223.000000
         mean       7.390583
         std        0.917130
         min        3.700000
         25%        6.900000
         50%        7.500000
         75%        8.000000
         max        9.500000
         Name: IMDb rating, dtype: float64
```

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

   Año de lanzamiento:

   - Media - esta medida nos dice que el año en el que tuvo más lanzamientos de serie fue en 2011.
   - Mediana - nos dice que el año en el que hay más lanzamientos a partir del 2016.
   - Desviación Estándar - Esta desviación estándar nos dice que hubo una gran variación en las fechas de lanzamiento lo que esperábamos realmente de este dato ya que las series siempre van a salir en fechas diferentes, pero si hubiese habido una desviación estándar baja, significa que muchas series salieron en muy poco tiempo.

   Número de temporadas disponibles:

   - Media - Como se pudo observar en la base de datos, se encuentra que en general, las series tienen aproximadamente 2.61 temporadas. No obstante, es evidente que las series no pueden estar a la mitad. Así pues, se puede decir que en promedio, las series tienen 3 temporadas disponibles.
   - Mediana - En el caso de la mediana, se puede decir que el primer 50% de los datos están más acumulados. Esto debido a que la mayor parte de las series tienen entre 1-2 temporadas.
   - Desviación Estándar - La desviación estándar nos dice que tan acoulados están nuestros datos, en este caso se obtuvo: 2.59.

Esto nos indica que realmente no hay mucha variación entre nuestros datos, ya que como veíamos con anterioridad, en realidad el primer 50% de nuestros datos se encuentran bastante acumulados.

Calificación en IMDb:

- Media - Este dato nos proporciona un gran acercamiento a la satisfacción que se les está dando a los clientes. Pues vemos cómo es que en satisfacción se tiene un promedio de 7.39. Se debe estudiar un poco más a detalle para ver qué es lo que está pasando con las series, para que de esta manera se pueda mejorar la calificación. Este dato nos puede proporcionar información sobre el tipo de contenido que el cliente está buscando, pues al saber las series que tienen mayor puntuación, se pueden recolectar las series que tengan contenido similar.
- Mediana - lo que nos dice que la mayoría de las series tiene una clasificación mayor a 7.5 y son muy pocas las que tienen valores menores a este
- Desviación Estándar - 0.9 este dato nos dice que la clasificación de las series está muy reñida ya que la gente tiene una opinión similar acerca de la series en esta plataforma por lo que podemos destacar que 7 no es tan mala clasificación pero queda algo por mejorar.

La siguiente tabla muestra los datos de todas las variables en conjunto como arriba se mostró previamente.

```
In [27]: base.describe()
```

Out[27]:

	S.no.	Year of release	No of seasons available	IMDb rating
count	393.000000	393.000000	393.000000	223.000000
mean	197.000000	2011.274809	2.608142	7.390583
std	113.593574	12.943787	2.592008	0.917130
min	1.000000	1926.000000	1.000000	3.700000
25%	99.000000	2011.000000	1.000000	6.900000
50%	197.000000	2016.000000	2.000000	7.500000
75%	295.000000	2018.000000	3.000000	8.000000
max	393.000000	2020.000000	20.000000	9.500000