Universidad del Valle de Guatemala
Data Preparation
Alexis Fernando Hengstenberg Chocooj

Proyecto Final

Algoritmo

Se utilizo el lenguaje de programación de Python, el cual es muy utilizado para realizar este tipo de tareas, a continuación, una explicación del algoritmo.

Se utilizaron las siguientes librerías:

```
import pandas as pd
import os
import datetime
import chardet
```

Las cuales son fundamentales para este trabajo.

- Pandas: librería utilizada para la manipulación y transformación de datos.
- Os: librería utilizada para leer los archivos en un directorio.
- Datetime: librería para trabajar con fechas.
- Chardet: se utilizo para detectar la codificación de los archivos de texto y poder abrirlos.

En el siguiente fragmento utilizamos la librería de chardet para leer que tipo de codificación tienen los archivos de texto y poder convertirlo y utilizarlo para nuestra investigación.

```
with open("./item.txt ", 'rb') as rawdata:
    result = chardet.detect(rawdata.read(100000))
result
```

A continuación, leemos el archivo item.txt y le asignamos encabezados, lo mismo sucede con el archivo de uses. Eliminamos los datos duplicados y eliminamos la columna que no utilizaremos en este caso no_used

```
item = pd.read_csv("./item.txt ", sep = "|", header=None, encoding='ISO-8859-1')
item = item.drop_duplicates()
item = item.set_axis(['itemid', 'movie_title', 'release_date', 'No_used',
                      "url","unknown","action","adventure","animation","childens","comedy","crime","documentary","drama",
                      "fantasi","film-noir","horror","musical","mystery","romance","sci-fi","thriller","war","western"], axis=1)
item = item.drop(['No_used'], axis=1)
```

A continuación, leemos todos los archivos CSV que contienen las calificaciones y los unimos para tener la información disponible. Le eliminamos los duplicados y las calificaciones que sean mayores a 6.

```
contenido = os.listdir('./movies_userdata')
combinado_csv = pd.concat([pd.read_csv("./movies_userdata/"+f) for f in contenido], ignore_index= True)
combinado_csv = combinado_csv.drop_duplicates()
combinado_csv.drop(combinado_csv[(combinado_csv['raiting'] > 5)].index, inplace=True)

combinado_csv['date'] = [datetime.datetime.fromtimestamp(s,datetime.timezone.utc) for s in combinado_csv['timestamp']]
#print(combinado_csv)
```

Luego asociamos los datos de las películas, los usuarios y las calificaciones para tenerlos completos.

```
pregunta4 = combinado_csv.merge(item, on='itemid', how='left')
```

Luego convertimos la fecha a una estructura mas accesible y legible para todos los usuarios.

```
combinado_csv['date'] = [datetime.datetime.fromtimestamp(s,datetime.timezone.utc) for s in combinado_csv['timestamp']]
#print(combinado_csv)
```

Para finalizar utilizamos distintas funciones que tiene la librería Python para ir respondiendo cómodamente las pregunta ya con la información limpia y preparada.

Para la pregunta 1 se agrupan por la película y luego se cuentan todas las calificaciones para saber cual fue la mas popular, luego se ordena de mayor a menor para tener las mas populares en la lista.

```
pregunta1 = combinado_csv.groupby(['itemid']).count().reset_index().sort_values('userid', ascending=False)
print(pregunta1)
```

Para la pregunta 2 se crea una variable nueva que es agrupada por la película y a su ves se va sumando el rating para encontrar las películas con las calificaciones mas altas.

```
pregunta2 = combinado_csv.groupby(['itemid']).agg(
                                {'raiting': 'sum',
                                }).reset_index()

print(pregunta2.sort_values('raiting', ascending=False).head(10))

movie1 = item.loc[item['itemid'] == 50]
```

Luego se intento realizar el mismo procedimiento, pero en vez de suma, quisimos obtener el promedio para tener una mejor idea, pero podemos observar que al existir películas con 1 sola calificación puede afectar en la investigación y se decidió no incluir estos datos.

```
pregunta2 = combinado_csv.groupby(['itemid']).agg(
                                {'raiting': 'mean',
                                }).reset_index()
```

Para la pregunta 3 al ser una variable de unos y ceros, se puede contar fácilmente la columna de cada una de las películas que se calificaron en cada género.

```
print('Unknown', item['unknown'].sum())
print('action', item['action'].sum())
print('adventure', item['adventure'].sum())
print('animation', item['animation'].sum())
print('childens', item['childens'].sum())
print('comedy', item['comedy'].sum())
print('crime', item['crime'].sum())
print('documentary', item['documentary'].sum())
print('drama', item['drama'].sum())
print('fantasi', item['fantasi'].sum())
print('film-noir', item['film-noir'].sum())
print('horror', item['horror'].sum())
print('musical', item['musical'].sum())
print('mystery', item['mystery'].sum())
print('romance', item['romance'].sum())
print('sci-fi', item['sci-fi'].sum())
print('thriller', item['thriller'].sum())
print('war', item['war'].sum())
print('western', item['western'].sum())
```

En la pregunta número 4 se realizó el mismo procedimiento anterior, pero en vez de contar o sumar la columna por cada genero se agruparon por género y se sumó el rating para saber que genero tenía mejores calificaciones.

```
unknown = pregunta4.groupby(['unknown']).agg({'raiting': 'sum',}).reset_index()
action = pregunta4.groupby(['action']).agg({'raiting': 'sum',}).reset_index()
adventure = pregunta4.groupby(['adventure']).agg({'raiting': 'sum',}).reset_index()
animation = pregunta4.groupby(['animation']).agg({'raiting': 'sum',}).reset_index()
childens = pregunta4.groupby(['childens']).agg({'raiting': 'sum',}).reset_index()
comedy = pregunta4.groupby(['comedy']).agg({'raiting': 'sum',}).reset_index()
crime = pregunta4.groupby(['crime']).agg({'raiting': 'sum',}).reset_index()
documentary = pregunta4.groupby(['documentary']).agg({'raiting': 'sum',}).reset_index()
drama = pregunta4.groupby(['drama']).agg({'raiting': 'sum',}).reset_index()
fantasi = pregunta4.groupby(['fantasi']).agg({'raiting': 'sum',}).reset_index()
film_noir = pregunta4.groupby(['film-noir']).agg({'raiting': 'sum',}).reset_index()
horror = pregunta4.groupby(['horror']).agg({'raiting': 'sum',}).reset_index()
musical = pregunta4.groupby(['musical']).agg({'raiting': 'sum',}).reset_index()
mystery = pregunta4.groupby(['mystery']).agg({'raiting': 'sum',}).reset_index()
romance = pregunta4.groupby(['romance']).agg({'raiting': 'sum',}).reset_index()
sci_fi = pregunta4.groupby(['sci-fi']).agg({'raiting': 'sum',}).reset_index()
thriller = pregunta4.groupby(['thriller']).agg({'raiting': 'sum',}).reset_index()
war = pregunta4.groupby(['war']).agg({'raiting': 'sum',}).reset_index()
western = pregunta4.groupby(['western']).agg({'raiting': 'sum',}).reset_index()
```

En la pregunta 5 solo se calculo el promedio de las edades de las personas que realizaron una calificación.

```
print('La media de edad de las personas es de:\n',users['age'].mean())
```

Para lograr responder la pregunta 6 se realizaron distintas gráficas, para ver el comportamiento de las personas y si tenia algo que ver la edad con la calificación que daban, para esto se utilizó la librería de matplot para generar graficas.

```
import matplotlib.pyplot as plt


scatter_plot = pregunta6.plot.scatter(x='age',y='raiting')
scatter_plot.plot()
plt.show()
```

Continuamos con la pregunta numero 7 y para aquí se eliminaron las calificaciones de personas que eran menores de 50 años para luego agrupar por película y ver que película tenía más calificaciones.

```
pregunta7.drop(pregunta7[(pregunta7['age'] < 50)].index, inplace=True)


resultado = pregunta7.groupby(['itemid']).count().reset_index().sort_values('userid', ascending=False)
print(resultado)
```

En la penúltima pregunta se utilizaron los datos combinados de usuarios, películas y calificaciones para ver cuantas calificaciones tenían las películas, pero estas utilizadas agrupadas por género.

```
print('Unknown', item['unknown'].sum())
print('action', item['action'].sum())
print('adventure', item['adventure'].sum())
print('animation', item['animation'].sum())
print('childens', item['childens'].sum())
print('comedy', item['comedy'].sum())
print('crime', item['crime'].sum())
print('documentary', item['documentary'].sum())
print('drama', item['drama'].sum())
print('fantasi', item['fantasi'].sum())
print('film-noir', item['film-noir'].sum())
print('horror', item['horror'].sum())
print('musical', item['musical'].sum())
print('mystery', item['mystery'].sum())
print('romance', item['romance'].sum())
print('sci-fi', item['sci-fi'].sum())
print('thriller', item['thriller'].sum())
print('war', item['war'].sum())
print('western', item['western'].sum())
resultado8 = pregunta8.groupby(['gender']).agg(
                                    {'unknown': 'sum',
                                     'action': 'sum',
                                     'adventure': 'sum',
                                     'animation': 'sum',
                                     'childens': 'sum',
                                     'comedy': 'sum',
                                     'crime': 'sum',
                                     'documentary': 'sum',
```

Por último, para la pregunta numero 9 se utilizo la columna de la fecha ya convertida con anterioridad para facilitarnos encontrar los fines de semana, con la librería de date time podemos identificar día de la semana es una fecha en específico, en este caso eliminamos los que días que no fueran sábados ni domingos, luego eliminamos las horas que estén fuera del horario de 8 de la mañana a 10 de la noche y agrupamos las calificaciones por película y sumamos todas para ver cuáles eran las más populares.

```
combinado_csv.drop(combinado_csv[(combinado_csv["dia"] < 5)].index, inplace=True)
combinado_csv.drop(combinado_csv[(combinado_csv["hora"] < 8)].index, inplace=True)
combinado_csv.drop(combinado_csv[(combinado_csv["hora"] > 20)].index, inplace=True)
combinado_csv.drop(combinado_csv[(combinado_csv["hora"] == 20) & (combinado_csv["minuto"] > 0)].index, inplace=True)
print(combinado_csv)
resultado9 = combinado_csv.groupby(['itemid']).agg(
                                {'raiting': 'sum',
                                }).reset_index()
```

Preguntas

- **¿Cuáles son las 3 películas más populares? Es decir, las que más calificaciones han recibido, sin importar la calificación otorgada.**

```
VG/Data Preparation/data.py"
      itemid  userid  raiting  timestamp  date
49        50     583      583        583   583
257      258     509      509        509   509
99       100     508      508        508   508
180      181     507      507        507   507
293      294     485      485        485   485
...      ...     ...      ...        ...   ...
1575    1576       1        1          1     1
1576    1577       1        1          1     1
1347    1348       1        1          1     1
1578    1579       1        1          1     1
1681    1682       1        1          1     1

[1682 rows x 5 columns]
```

Las películas mas populares se encuentran con el código 50, 258, 100

```
Star Wars (1977)
movie_title, dtype: object
 Contact (1997)
movie_title, dtype: object
Fargo (1996)
movie_title, dtype: object
 Return of the Jedi (1983)
movie_title, dtype: object
```

Las películas mas populares por nombre son:

- Star Wars (1977)
- Contact (1997)
- Fargo (1996)
- Return of the Jedi (1983)

En ese orden.

- **¿Cuáles son las 10 películas mejor calificadas por los usuarios?**
    - Top 10 de pelicumas con mejores calificaciones sumadas

itemid	raiting
50	2541
100	2111
181	2032
258	1936
174	1786
127	1769
286	1759
1	1753
98	1673
288	1645

    - **Son las siguientes :**
    - Star Wars (1977)
    - Fargo (1996)
    - Return of the Jedi (1983)
    - Contact (1997)
    - Raiders of the Lost Ark (1981)
    - Godfather, The (1972)
    - English Patient, The (1996)
    - Toy Story (1995)
    - Silence of the Lambs, The (1991)
    - Scream (1996)

```
Star Wars (1977)
movie_title, dtype: object
Fargo (1996)
movie_title, dtype: object
 Return of the Jedi (1983)
movie_title, dtype: object
 Contact (1997)
movie_title, dtype: object
 Raiders of the Lost Ark (1981)
movie_title, dtype: object
 Godfather, The (1972)
movie_title, dtype: object
 English Patient, The (1996)
movie_title, dtype: object
Toy Story (1995)
movie_title, dtype: object
Silence of the Lambs, The (1991)
movie_title, dtype: object
 Scream (1996)
movie_title, dtype: object
```

Para realizar un promedio de las calificaciones se obtiene una mayor distorsión ya que las películas con menos calificaciones y con mejor calificación pueden llegar a salir primero como en el siguiente ejemplo

	itemid	raiting
813	814	5.0
1598	1599	5.0
1200	1201	5.0
1121	1122	5.0
1652	1653	5.0
1292	1293	5.0
1499	1500	5.0
1188	1189	5.0
1535	1536	5.0
1466	1467	5.0

```
 Great Day in Harlem, A (1994)
movie_title, dtype: object
  Someone Else's America (1995)
movie_title, dtype: object
  Marlene Dietrich: Shadow and Light (1996)
movie_title, dtype: object
  They Made Me a Criminal (1939)
movie_title, dtype: object
  Entertaining Angels: The Dorothy Day Story (1996)
movie_title, dtype: object
  Star Kid (1997)
movie_title, dtype: object
  Santa with Muscles (1996)
movie_title, dtype: object
  Prefontaine (1997)
movie_title, dtype: object
  Aiqing wansui (1994)
movie_title, dtype: object
  Saint of Fort Washington, The (1993)
movie_title, dtype: object
```

Con lo cual no puede ser fiable confiarnos del promedio. A pesar de que mientras más calificaciones tenga una película puede beneficiarse con la suma de la calificación es mucho mas confiable el primer análisis.

- **¿Cuántas películas hay de cada género? Si una película pertenece a más de un género, puede contarse más de una vez.**
  En total de películas de cada genero son las siguientes:

```
------------------------PREGUNTA 3---------------------------
action 251
adventure 135
animation 42
childens 122
comedy 505
crime 109
documentary 50
drama 725
fantasi 22
film-noir 24
horror 92
musical 56
mystery 61
romance 247
sci-fi 101
thriller 251
war 71
western 27
```

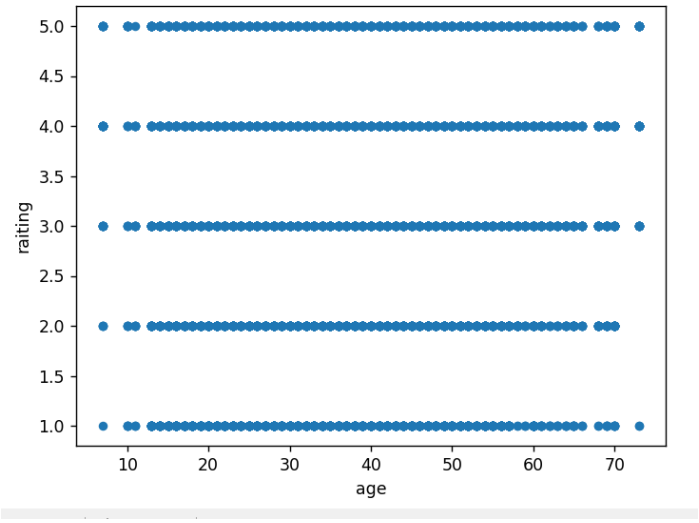
- **¿Qué género de películas tiene las calificaciones más altas?**
  El género que tiene calificación más alta es drama.

```
   ---------------------
      drama  raiting
  1       1   147108
      comedy  raiting
  1       1   101252
      action  raiting
  1       1    89056
      thriller  raiting
  1         1    76749
      romance  raiting
  1         1    70482
      adventure  raiting
  1           1    48184
      sci-fi  raiting
  1       1    45328
      war  raiting
  1     1    35861
      crime  raiting
  1       1    29258
```

- **¿Cuál es la edad media de todos los usuarios que calificaron las películas?**
    - La edad media de todas las personas que generaron una calificación es de 34 aproximadamente

```
----------------------PREGUNTA 5-----------
La media de edad de las personas es de:
 34.05196182396607
```

- **¿Existe alguna relación entre la edad del usuario y la calificación que otorgó a la película?**
    - No existe ninguna relación entre la edad y la calificación esto se puede ver en la siguiente grafica donde nos damos cuenta de que la dispersión es igual para todas las edades.



- **¿Cuáles son las películas más populares entre los usuarios mayores de 50 años?**
    - Las películas mas populares para los usuarios mayores de 50 años son:

```
rows x 9 columns]
 English Patient, The (1996)
movie_title, dtype: object
Fargo (1996)
movie_title, dtype: object
 Air Force One (1997)
movie_title, dtype: object
Star Wars (1977)
movie_title, dtype: object
```

    - English Patient.
    - Fargo
    - Air Force One
    - Star Wars

- **¿Hay algún género de película que sea más popular entre los usuarios mujeres que entre los usuarios hombres?**

gender	unknown	action	adventure	animation	childens	comedy	crime	documentary	drama	fantasi	film-noir	horror	musical	mystery	romance	sci-fi	thriller	war	western
F	2	5442	3141	995	2232	8068	1794	187	11008	363	385	1197	1442	1314	5858	2629	5086	2189	371
M	8	20147	10612	2610	4950	21764	6261	571	28887	989	1348	4120	3512	3931	13603	10101	16786	7209	1483

Podemos ver tenemos mayores calificaciones de hombres que de mujeres, y que en la mayoría de las películas los hombres han dado mas calificaciones que las mujeres, lo que nos dice que los hombres ven más películas. Y que en géneros como acción aventura si existe una diferencia visible, así como en algunos otros.

- **¿Cuál es la película mejor calificada durante fines de semana, en horario de 8 a 10 PM?**
  La película mejor calificada durante los fines de semana es

```
itemid  raiting
    50      210
   100      187
     7      185
    56      171
   258      171
   174      167
   181      166
   121      165
   117      162
    98      153
 Star Wars (1977)
: movie_title, dtype: object
 Fargo (1996)
: movie_title, dtype: object
Twelve Monkeys (1995)
```

  La película mejor calificada en horarios del fin de semana entre un horario de 8 de la mañana a 10 de la noche son:
  - Star Wars.
  - Fargo.
  - Twelve Monkeys.