Práctica 2: ¿Cómo realizar la limpieza y análisis de datos?

Descripción del dataset

El dataset con el que vamos a trabajar es *puntuacionesFA.csv.* Este contiene los datos del usuario de FilmAffinity 968973. Se compone de 1941 filas, las cuales corresponden a las películas valoradas por el usuario y 14 columnas que reflejan:

1. **titulo: Es el título de la película dado para el mercado español.**
2. **año: El año de estreno de la película.**
3. **duración: La duración en minutos del film.**
4. **pais: País donde se ha producido el metraje.**
5. **directores: Lista de director/es que realizaron la película.**
6. **guionistas: Lista de guionista/s que guionizaron la película.**
7. **musicos: Lista de musico/s implicado/s en la banda sonora del metraje.**
8. **fotografos: Lista de fotógrafo/s que participaron en la película.**
9. **actores: Lista de intérprete/s que actúan en la película.**
10. **productores: Lista de compañía/s que financian el metraje.**
11. **generos: Lista de genero/s entre los que se clasifica el metraje.**
12. **nota: La nota media de todos los usuarios de FilmAffinity que votaron el film.**
13. **votos: Número de votos que ha recibido la película por parte de los usuarios.**
14. **votacion: Votación del usuario del perfil del dataset.**

Este dataset es el obtenido en la práctica de web scraping anterior. El análisis de este juego de datos pretende incorporar nuevas herramientas de análisis que la plataforma de FilmAffinity no suple, para realizar un estudio más exhaustivo de los gustos o preferencias del usuario.

Integración y selección

Para este dataset no hemos considerado necesario añadir datos adicionales pero si crear varios subsets para el análisis posterior.

Este conjunto de datos contiene varias columnas que acumulan datos en formato de lista, las cuales hemos transformado mediante una función propia que se ejecuta con un archivo 2_extract_subsets.py. Este ejecutable permite añadir el nombre de cualquier columna del dataset inicial para crear el archivo csv que se requiera.

```python
import argparse
import pandas as pd


# Función para desglosar cualquier columna en varias filas
def desglosar_lista(row,column):
    transformada = row[column].strip("[]").split(", ")
    titulo = row["titulo"]
    pais = row["pais"]
    votos = row["votos"]
    votacion = row["votacion"]
    nota_media = row["nota"]
    desglosado = []
    for i in transformada:
        desglosado.append([i.strip("'"), titulo, pais, votos, nota_media, votacion])
    return desglosado


if __name__ == "__main__":
    parser = argparse.ArgumentParser(description='decompose puntuacionesFA in subsets')
    parser.add_argument("-c",
                        "--column",
                        type = str,
                        help="base column to decompose")
    args = parser.parse_args()
    column = args.column

    # Cargamos el dataframe original
    df = pd.read_csv('../datasets/puntuacionesFA.csv')

    # Aplicar la función a cada fila del DataFrame para desglosar la columna especificada
    filas_desglosadas = df.apply(lambda row: desglosar_lista(row, column), axis=1).explode()

    # Crear un nuevo DataFrame con las filas desglosadas
    new_df = pd.DataFrame(filas_desglosadas.tolist(), columns=[column, "titulo", "pais", "votos", "votacion", "nota"])
    new_df.to_csv("../datasets/" + column + "_subset.csv", index=False)
```

```
python 2_extract_subsets.py -c directores
```

Limpieza de los datos

Hemos detectado presencia de valores nulos en las columnas de duración, nota y votos.

```
titulo         0
año            0
duracion       2
pais           0
directores     0
guionistas     0
musicos        0
fotografos     0
actores        0
productores    0
generos        0
nota           3
votos          3
votacion       0
```

Los nulos detectados en las columnas de nota y votos no las podemos sustituir por valores artificiales por lo que hemos decidido eliminar estas filas, debido a que no representan un alto porcentaje del total de registros.

En cuanto a los valores nulos en duración los hemos sustituido por una media del resto de valores de la columna.

Al realizar el análisis de valores extremos hemos detectado la presencia de estos en las columnas año, duración, nota, votos y votación. Como explicamos a continuación debido a la naturaleza de los datos estos valores extremos son normales.

-año: Los datos extremos se deben a que existen películas muy antiguas y el usuario ha visualizado aquellas más recientes.

-duracion: Se encuentran estos valores dispares ya que Filmaffinity contiene los datos tanto de contometrajes como de documentales de larga duración.

-nota: Las valoraciones más bajas son las que causan estos valores atípicos, pero no se salen del rango 0-10.

-votos: Solo refleja la cantidad de votaciones de los usuarios de la plataforma.

-votacion: Al igual que la nota no se sale del rango 0-10.

Análisis de los datos

El análisis de los datos se puede aplicar a los siguientes subsets:

-directores_subset.csv: Contiene en desglose de los valores de la columna directores con sus valores correspondientes de titulo, votos, nota y votación.

-generos_subset.csv: Contiene en desglose de los valores de la columna generos con sus valores correspondientes de titulo, votos, nota y votación.

-guionistas_subset.csv: Contiene en desglose de los valores de la columna guionistas con sus valores correspondientes de titulo, votos, nota y votación.

Debido a que sería un análisis repetitivo vamos a aplicar el análisis al subset directores_subset.csv.

:[43]:

	directores	titulo	votos	votacion
0	Todd Phillips	Joker	69408.0	10
1	Damien Chazelle	La ciudad de las estrellas (La La Land)	58319.0	10
2	Christopher Cantwell	Halt and Catch Fire (Serie de TV)	4517.0	10
3	Christopher C. Rogers	Halt and Catch Fire (Serie de TV)	4517.0	10
4	Karyn Kusama	Halt and Catch Fire (Serie de TV)	4517.0	10
...	...	...	...	...
3174	Mark Mylod	Ali G anda suelto	36100.0	1
3175	David Lynch	Darkened Room (C)	641.0	1
3176	Juan Muñoz	¡Ja me maaten...!	4379.0	1
3177	Kinji Fukasaku	Battle Royale	35461.0	1
3178	Marv Newland	Bambi Meets Godzilla (C)	5489.0	1

3179 rows × 4 columns

CORRELACIÓN ENTRE EL NÚMERO DE VOTOS Y LA NOTA MEDIA DE LAS PELÍCULAS:

Analizamos si el hecho de que una película sea vista por más gente implica mejor puntuación. Hemos descubierto que existe asociación positiva entre los dos hechos pero no se puede demostrar la razón:

-Los usuarios visualizan más una película según su buena nota.

-Se condiciona más la valoración de una película según su popularidad en estreno.

CORRELACIÓN ENTRE LA NOTA MEDIA DE LA PELICULA Y LA NOTA MEDIA DEL USUARIO:

Se busca comprobar si el voto del usuario es similar al voto de la comunidad. Como conclusión hemos obtenido que el usuario vota de manera similar al resto de usuarios de la plataforma.

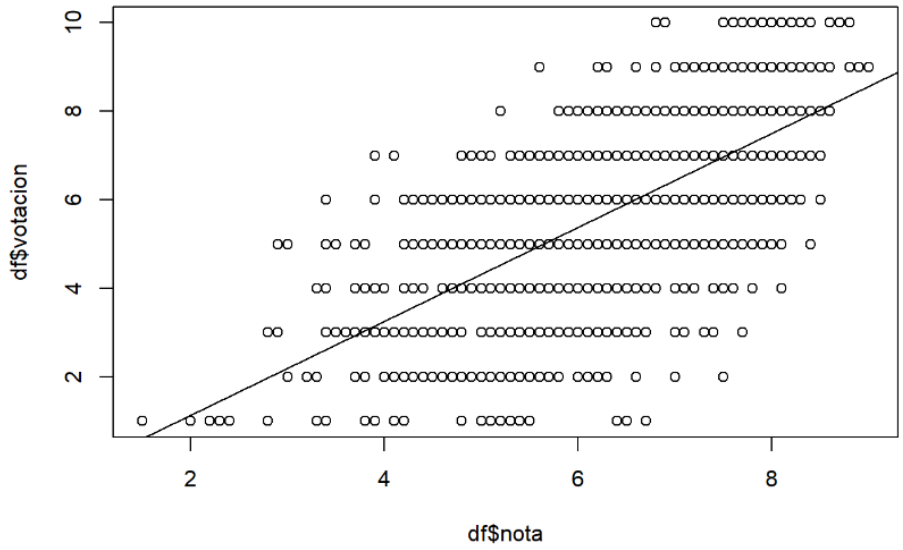
En el análisis estadístico en primer lugar hemos analizado el dataset principal para resolver preguntas sobre como se comportan las votaciones en general y en relación al usuario.

CUANTITATIVO VS CUANTITATIVO

Hemos analizado la correlación entre el número de votos que tiene una película y la nota media que obtiene, siendo esta correlación positiva y significativa. Esto quiere decir que a medida que aumenta el número de votos aumenta la nota media, y viceversa. No podemos saber cual de las dos variables condiciona la otra, pero es esperable pensar que la nota media positiva haga que mas gente vaya a ver la película.

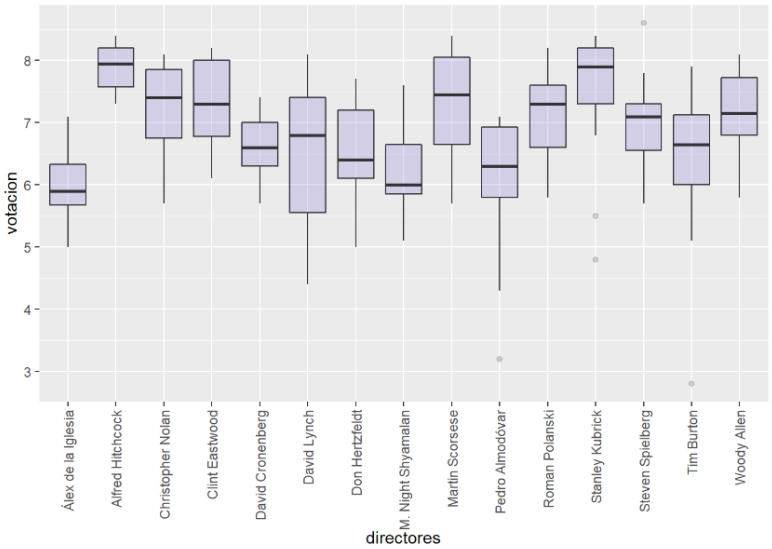
Por otro lado, hemos analizado la correlación entre la nota media de la película y la nota media del usuario. Esta correlación nos ha salido positiva y significativa, más significativa que la previa. Esto quiere decir que a medida que la nota media es más alta la votación del usuario también lo es. Esto es útil para que el usuario sepa que se puede fiar en cierta medida de la nota media de FilmAffinity, ya que refleja sus gustos.

Esto lo hemos representado mediante un gráfico



CUANTITATIVO VS CUALITATIVO

En este apartado hemos analizado el subset de directores. Hemos escogido los directores con más de 10 películas votadas, para tener suficiente N. Hemos representado tanto el número de películas votadas por estos directores como la votación de dichas películas en un boxplot.



Aquí se observan cuales son los directores con mejor mediana: Stanley Kubrick y Alfred Hitchcock.

Hemos querido estudiar si hay diferencia entre ellos. Para ello hemos comprobado si cumplían la normalidad. Lo cual no hacían. Tampoco teníamos suficiente número para asumir la teoría del límite central. Hemos calculado la homocedastidad solo como parte del proceso, aunque realmente no haría falta ya que sin normalidad no podemos aplicar un test paramétrico (t-student), por lo tanto, hemos aplicado el test no paramétrico para comparar dos grupos (Wilcoxon test). Este test nos ha demostrado que no existen diferencias entre la valoración del usuario entre ambos directores.

A continuación, hemos decidido comprobar las diferencias entre TODOS los directores con más de 10 películas votadas. En este caso tampoco se cumplía la normalidad de los datos, por lo que en vez de usar el test paramétrico para compara medias entre más de dos grupos (ANOVA), hemos utilizado un test no paramétrico (Kruskall-Wallis). En este test hemos visto que efectivamente existían diferencias significativas. Sin embargo, este resultado no nos dice entre que directores existen las diferencias, por lo cual hemos utilizado un test post-hoc con corrección de Bonferroni, que ajusta el p-valor para el número de comparaciones que se realizan (Pairwise Wilcoxon test con corrección de Bonferroni). Con este test hemos visto entre que directores existían diferencias significativas en la valoración del usuario.

<u>CUANTITATIVO VS CUANTITATIVO</u>

Por último, hemos querido estudiar si dentro de los directores con más de 10 votaciones había asociaciones entre los mismo y el país. Para ello en primer lugar hemos hecho una tabla de contingencia con proporciones, que mostraban la proporción de películas que cada director había realizado en cada país.

Posteriormente hemos querido comprobar si había asociaciones entre directores y país. Para

ello hemos utilizado el test de Chi cuadrado, que ha mostrado que efectivamente existían diferencias. El problema de nuevo, es que este test nos habla de que existen diferencias globalmente, pero no entre que grupos existen. Para comprobar esto último hemos aplicado un test de chi-cuadrado post-hoc con corrección de Bonferroni, que hace una comparación entre cada uno de los directores con cada país para mostrarnos si existen asociación.

Resolución del problema

Estos análisis nos han servido para saber que el usuario puede fiarse de las notas de FilmAffinity para elegir futuras películas, ya que guarda mucha relación con ellas. También sabe que directores son los mejor valorados, para buscar futuras películas con ellos. Por otro lado, este tipo de análisis se podría haber hecho también con países, con duración. Y hay varios subsets que hemos creado que se podrían haber analizado igual que este, para comprobar cuales son los guionistas favoritos, géneros favoritos, fotógrafos favoritos etc.

Contribuciones

Contribuciones	Firma
Investigación previa	MR,CR
Redacción de las respuestas	MR,CR
Desarrollo del código	MR,CR
Participación en el vídeo	MR,CR