Sergi Mas Rodriguez – 47793952K

Tipologia i cicle de vida de les dades - PRAC 2

Consideracions prèvies:

- Tal i com vaig detallar en l'anterior entrega, el motiu pel qual he decidit fer les dues pràctiques individualment enlloc per parelles tal i com es demanava des del professorat de l'assignatura és la meva reduïda i volàtil disponibilitat horària.
- Per altra banda i seguint el mateix esquema que en l'anterior pràctica, en aquest document intentaré anar donant resposta als diferents apartats que es plantegen a la secció de *Descripció de la pràctica a realitzar*. Addicionalment, s'ha realitzat un *README.md* a GitHub on es detallen les diferents parts del repositori.

Així doncs, comencem a detallar els diferents apartats que se'ns demana per a la realització d'aquesta pràctica:

1. Descripció del dataset:

El nom del data set és *db_FA.csv* i conté els següents atributs:

- Id
- Títol
- Any
- Duració
- País
- Director/s
- Guionista/es
- Música
- Fotografia
- Productora
- Repartiment
- Gènere
- Nota
- Votacions
- Pàgina web oficial

Els registres que hi haurà en aquest data set corresponen a les primeres 500 pel·lícules que apareixen com a resultat de la cerca feta per cada país que podíem consultar (incloent la categoria *ALL*) i que han estat filmades entre l'any 2015 i 2019. Teòricament el nostre data set hauria de contenir 4500 registres (500 per cada país disponible) però m'he trobat que en certs moments de la descàrrega creava l'error de '*Too many requests*' i em baixava menys pel·lícules de les que m'havia de baixar. Addicionalment, com que també incloem la categoria de país 'ALL' és molt possible que s'hagin produït algunes duplicitats.

Amb aquest data set es pretén analitzar les següents preguntes:

1. Existeix una relació entre major nombre de votacions i major qualitat?
2. Existeix un país que produeixi pel·lícules de major qualitat segons els usuaris de FA?
3. Existeix una categoria (gènere de pel·lícula) amb millor puntuació que les altres categories pels usuaris de FA?
4. Els documentals estan més ben valorats que les sèries de TV?
5. Els curtmetratges tenen millors votacions que els llargmetratges?

Altres preguntes que haguessin estat interessants de respondre però que, malauradament, el nostre data set no permet respondre perquè no disposa de la informació seria si les pel·lícules amb premis de crítica professional tenen millor puntuació que pel·lícules sense premis. Moltes vegades s'ha dit que el gust dels crítics cinèfils és molt diferent al gust del gran públic i aquesta hagués estat una bona pregunta a respondre però que, malauradament, no es podrà respondre amb les dades disponibles.

2. Integració i selecció de les dades d'interès:

Totes les preguntes tenen a veure amb la puntuació mitja que han atorgat els usuaris de FA a les diferents pel·lícules. Així, podem establir que aquest és l'atribut clau del nostre data set. Per altra banda, eliminarem aquells atributs que no necessitem per així quedar-nos amb un conjunt més reduït i no consumir tanta memòria.

En aquest sentit, el data set resultant tindrà només els següents atributs:

- Any
- Duració
- País
- Gènere
- Nota
- Votacions

3. Pre - processament de les dades:

En un primer anàlisi de l'atribut any em vaig trobar que hi havia registres mal entrats ja que corresponien a anys que no havia descarregat. Això es deu que no devia borrar el data set que vaig crear de proves per a la primera pràctica i se'm van emmagatzemar dades que no pertocaven:

```
In [145]: # Comrpovació dels anys que hi ha en la nostra base
          print(FA['Any'].unique())
          print(FA.shape)

          [2019. 2000. 1999. 1998. 1997. 2018. 2017.]
          (5171, 15)
```

Vam fer un simple filtre de les dades per aconseguir tenir el nostre conjunt a treballar:

```
FA2 = FA[(FA['Any']==2017.) | (FA['Any']==2018.) | (FA['Any']==2019.)]
print(FA2['Any'].unique())
print(FA2.shape)

[2019. 2018. 2017.]
(4744, 15)
```

Per altra banda, com que vam seleccionar la possibilitat 'ALL', la probabilitat que hi hagués duplicacions en el nostre data set era molt alta així que es va realitzar un procés d'eliminació i reindexat del dataframe per tal d'eliminar aquestes problemàtiques:

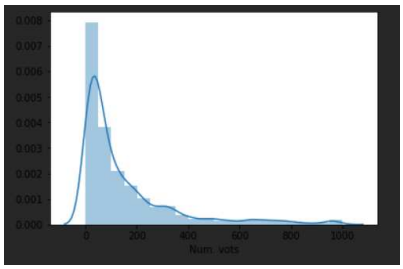
```
# Eliminació de registres repetits
FA3 = FA2.drop_duplicates(keep=False)
FA3 = FA3.reset_index(drop=True)
print(FA3.shape)

(3997, 15)
```

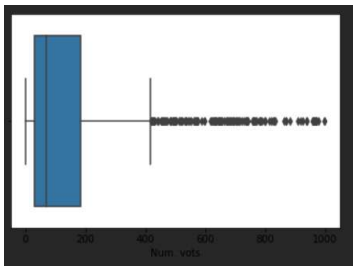
Per altra banda, també teníem la problemàtica de valors NA's en alguns registres i, per no alterar els testos estadístics que es volen aplicar més endavant, vaig optar per la seva eliminació ja que la imputació dels NA's a un valor específic ens hagués pogut conduir a conclusions errònies.

```
In [149]: FA_clean.isna().sum()

Any               0
Duracio (min)     380
Pais              0
Gènere            0
Nota              2114
Num. vots         2114
dtype: int64
```

```
In [150]: FA_clean = FA_clean.dropna()
          FA_clean = FA_clean.reset_index(drop=True)
          FA_clean.shape

(1774, 6)
```

L'únic atribut que podia contenir valors extrems era el corresponent a Nombre de votacions i s'ha detectat que la distribució d'aquest atribut té una *skewness* molt positiva:





Per aquests registres no s'ha realitzat cap transformació perquè corresponen a les pel·lícules més taquilleres i que han rebut més votacions dels últims anys.

Addicionalment, m'he trobat que els valors dels atributs Gènere, Num. Vots i Nota s'havien descarregat erròniament:

- En el cas de Gènere, prenia un string de la llista que generava el programa de web-scraping.
- En el cas de Num.Vots i Nota els nombres van ser emmagatzemats amb el format espanyol i no anglosaxó (és a dir, XX,XX quan hauria de ser XX.XX).

En aquest punt, m'haig de fer una autocrítica cap al meu propi programa perquè el repte que va suposar la transformació dels valors de Gènere va ser bastant gran i vaig perdre moltes hores intentant solucionar de forma senzilla la problemàtica que em vaig trobar. La solució proposada en el programa, però em va semblar que era la més elegant i fàcil d'aconseguir.

La resta de valors dels altres atributs no vaig haver de fet cap altre transformació.

4. Anàlisi dels atributs:

En el notebook de Jupyter, adjunto les diferents gràfiques de distribució dels valors, tant pels atributs continus com categòrics. Pels atributs categòrics s'adjunta un *countplot* per a cada categoria mentre que pels continus un *distplot*.

No he cregut necessari adjuntar en aquest document els diferents gràfics perquè seria duplicar els resultats i no hi veig el sentit.

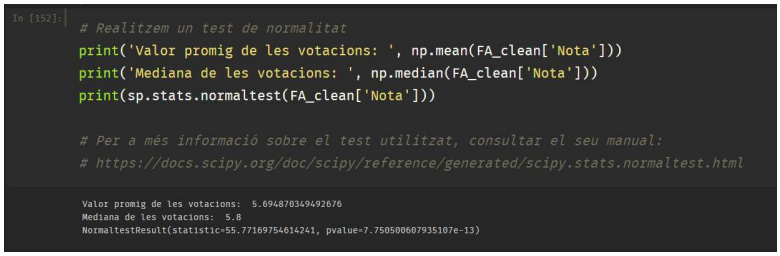
5. Resolució de problemes:

Com s'ha dit en el primer apartat, les preguntes que es volen respondre amb aquest data set són:

1. Existeix una relació entre major nombre de votacions i major qualitat?
2. Existeix un país que produeixi pel·lícules de major qualitat segons els usuaris de FA?
3. Existeix una categoria (gènere de pel·lícula) amb millor puntuació que les altres categories pels usuaris de FA?
4. Els documentals estan més ben valorats que les sèries de TV?
5. Els curtmetratges tenen millors votacions que els llargmetratges?

Per respondre a la primera pregunta utilitzo una senzilla OLS, tot graficant el seu resultat i, addicionalment, mostro els valors estadístics de la regressió. Com es pot veure en el notebook, la regressió és quasi una constant (el valor del regressor dependent és molt pròxim a zero) i, per tant, la pregunta plantejada resulta ser falsa. No existeix cap tipus de relació entre nombre de votacions i valoració de la pel·lícula.

Per respondre a les altres preguntes, comprovem si la distribució dels valors de l'atribut 'Nota' segueix una normal:



Com que el p-valor és molt pròxim a zero, podem rebutjar la hipòtesis nul·la i, per tant, la distribució no segueix una normal. Això implica que per tots els testos estadístics haguem d'utilitzar test no-paramètrics (de fet, utilitzarem el test de https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test) perquè el t-test sempre suposa normalitat de les distribucions.

Així, per la resta de preguntes plantejades tenim el següent:

2. El país que obté una millor valoració és Japó i la comparativa de mitjanes resulta significativa (*mu_jap* > *mu_njap*, essent *njap* la distribució de les notes dels països ex-Japó) .

3. El gènere que obté millor valoració resulta ser el Western però, en aquest cas, ens trobem que la diferència no resulta ser estadísticament significativa i, per tant, no podem rebutjar la hipòtesi nul·la. En aquesta pregunta s'ha de dir que el gènere que té una major valoració és el Documental però s'ha omès de l'anàlisi per no ser un gènere com a tal sinó una tipologia de filmació.

4. Els documentals resulten estar més ben valorats que les sèries de TV i, aquesta diferència, resulta ser significativa.

5. Noto que hem definit com a curtmetratge qualsevol film d'una durada igual o inferior als 30 minuts i com a llargmetratge la resta de filmacions. Els curtmetratges resulten estar més ben valorats que els llargmetratges i, aquesta diferència, resulta ser significativa.

6. Codi:

El codi s'ha realitzat a partir d'un Jupyter Notebook i en el GitHub s'inclou el codi així com la seva execució i sortida en HTML per la seva fàcil visualització. He cregut adient utilitzar un fiytxer *.ipynb enlloc del *.py estàndar perquè amb el primer et permet la combinació de text normal amb codi i la visualització de les gràfiques i l'estudi de la base resulta més fàcil i entenedora.

L'enllaç de GitHub és: https://github.com/Sergi-MR/prac2_tipologia_UOC