Pràctica 2. Tipologia i cicle de vida de les dades

Efecte de les Característiques de les Pel·lícules a la seva Durada

Silvia Galan Martínez i David Roche Valles

Github: https://github.com/sgalan/Practica2-Neteja

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

L'objectiu d'aquesta pràctica és estudiar la durada de les pel·lícules segons les seves característiques. Actualment, degut als canvis socials en les persones que consumeixen cinema, s'ha detectat que la durada de les filmacions (sigui en sèries o pel·lícules) afecta molt a l'èxit o consum d'aquest tipus de producte. Per tant, és d'interès qualsevol coneixement que es pugui extreure sobre la durada i la seva relació amb altres característiques dels films.

La principal idea, per tant, d'aquest estudi, és veure quines característiques de les pel·lícules poden tenir impacte sobre la seva durada i veure fins a quin punt a partir d'elles es pot predir la durada d'un film en concret.

Les etapes que es cobriran seran la selecció i neteja de les dades, l'anàlisi i la representació dels resultats obtinguts.

Les dades de les quals s'obté la informació necessària per assolir l'objectiu esmentat són dades de la plataforma Kaggle (https://www.kaggle.com/), en concret del data set "TMDB 5000 Movie

Dataset" (https://www.kaggle.com/tmdb/tmdb-movie-metadata) on es recullen 5.043 pel·lícules (files) i 28 característiques (columnes). Les variables que trobem són les següents:

- **Color**: informa si la pel·lícula és en blanc i negre o color.
- **Director_name**: el nom del director/a.
- **Num_critic_for_reviews**: Nombre de crítiques.
- **Duration**: duració de la pel·lícula.
- **Director_facebook_likes**: Quantitat de likes al facebook que té el director/a.
- **Actor_1_facebook_likes**: Quantitat de likes al facebook que té l'actor/actriu principal.
- **Actor_2_facebook_likes**: Quantitat de likes al facebook que té l'actor/actriu secundari.
- **Actor_3_facebook_likes**: Quantitat de likes al facebook que té l'actor/actriu terciari.
- **Actor_1_name**: Nom de l'actor/actriu principal.
- **Actor_2_name**: Nom de l'actor/actriu secundari.
- **Actor_3_name**: Nom de l'actor/actriu terciari.
- **Gross**: recaptació total bruta feta per la pel·lícula.
- **Genres**: Llista de gèneres per qualificar la pel·lícula.
- **Movie_title**: Títol de la pel·lícula.
- **Num_voted_users**: Nombre d'usuaris que han votat o valorat.
- **Cast_total_facebook_likes**: Quantitat total de likes al facebook que té tot l'equip que ha participat a la pel·lícula.
- **Facenumber_in_poster**: Número de rostres que apareixen en el pòster.
- **Plot_keywords**: Paraules clau per definir la pel·lícula.
- **Movie_imdb_link**: Link de la pel·lícula a IMDB.
- **Num_user_for_reviews**: Nombre de valoracions per usuaris.
- **Language**: Idioma (English, Spanish, Mandarin, etc.)
- **Country**: País.
- **Content_rating**: Valoració total del contingut.
- **Budget**: Pressupost per realitzar la pel·lícula.
- **Title_year**: Any que s'estrena el film.
- **Imdb_score**: Valoració del IMDB.
- **Aspect_ratio**: Rati o mida de la presentació del film al cinema (1.78, 2.35, etc.).
- **Movie_facebook_likes**: Quantitat de likes al facebook que té la pel·lícula.

Com l'objectiu final és estudiar i predir la durada d'un film a partir de les seves característiques informades en aquesta base de dades, les tècniques que s'utilitzaran a part de la neteja de dades, seran correlacions i contrastos per veure quines variables s'associen a la durada dels films i un model de regressió lineal múltiple per determinar com expliquen i el poder de predicció que tenen les característiques dels films sobre la seva durada.

Aquest anàlisis té rellevància en el sector cinematogràfic, ja que informa sobre la importància de la duració de les pel·lícules i la seva acceptació o valoració per part dels espectadors.

2. Integració i selecció de les dades d'interès a analitzar.

Les variables que seleccionem i mantenim per aquest anàlisi són les rellevants pel nostre objectiu: color, num_critic_for_reviews, duration, director_facebook_likes, actor_3_facebook_likes, actor_2_facebook_likes, actor_1_facebook_likes, gross, genres, num_voted_users, cast_total_facebook_likes, num_user_for_reviews, language, country, budget, title_year, imdb_score, aspect_ratio, movie_facebook_likes.

3. Neteja de les dades.

Primer de tot visualitzarem el dataset en format xlsx utilitzant el paquet *pandas (df.head)*.

Out[3]:

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_2_name	actor_1_facebook_likes	gross
0	Color	James Cameron	723.0	178.0	0.0	855.0	Joel David Moore	1000.0	760505847.0
1	Color	Gore Verbinski	302.0	169.0	563.0	1000.0	Orlando Bloom	40000.0	309404152.0
2	Color	Sam Mendes	602.0	148.0	0.0	161.0	Rory Kinnear	11000.0	200074175.0
3	Color	Christopher Nolan	813.0	164.0	22000.0	23000.0	Christian Bale	27000.0	448130642.0
4	NaN	Doug Walker	NaN	NaN	131.0	NaN	Rob Walker	131.0	NaN

Després identifiquem el tipus de variable de cada columna, per fer-ho usarem *df.dtypes* comprovant que estan en el format correcte*:*

```
Out[4]: color                      object
        director_name              object
        num_critic_for_reviews    float64
        duration                  float64
        director_facebook_likes   float64
        actor_3_facebook_likes    float64
        actor_2_name               object
        actor_1_facebook_likes    float64
        gross                     float64
        genres                     object
        actor_1_name               object
        movie_title                object
        num_voted_users             int64
        cast_total_facebook_likes   int64
        actor_3_name               object
        facenumber_in_poster      float64
        plot_keywords              object
        movie_imdb_link            object
        num_user_for_reviews      float64
        language                   object
        country                    object
        content_rating             object
        budget                    float64
        title_year                float64
        actor_2_facebook_likes    float64
        imdb_score                float64
        aspect_ratio              float64
        movie_facebook_likes        int64
        dtype: object
```

3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Identificació:

Els elements buits es troben el dataset com a *Na.* Observem el número d'elements buits per variable:

```
color 19
num_critic_for_reviews 50
duration 15
director_facebook_likes 104
actor_3_facebook_likes 23
actor_2_facebook_likes 13
actor_1_facebook_likes 7
gross 884
genres 0
num_voted_users 0
cast_total_facebook_likes 0
num_user_for_reviews 21
language 12
country 5
budget 492
title_year 108
imdb_score 0
aspect_ratio 329
movie_facebook_likes 0
```

Tractament:

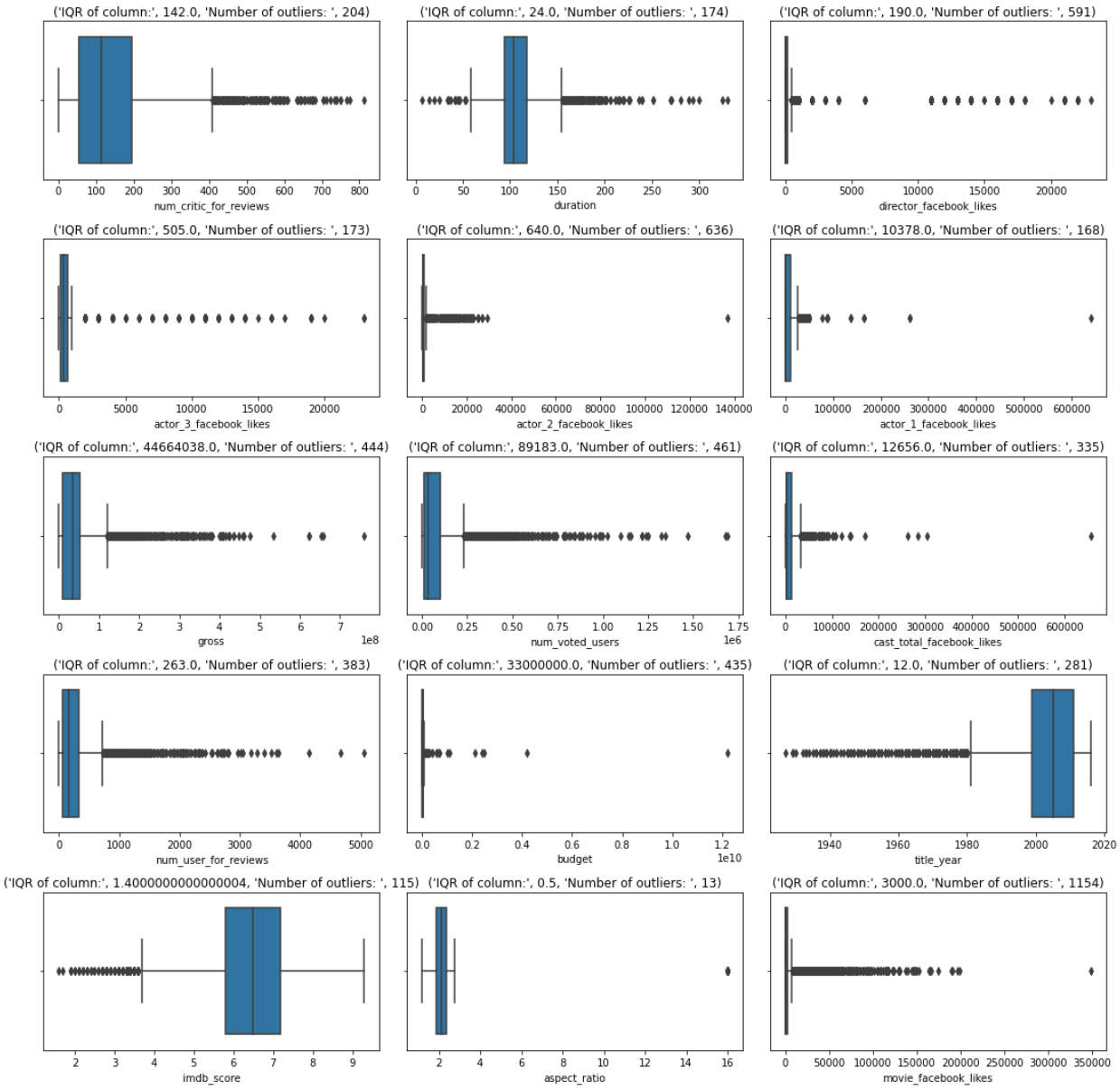
Els elements buits de les variables categòriques (genres, language, country, color i title_year), les eliminem, obtenint un dataset de 4.913 pel·lícules.
Pel que fa als elements buits que es troben a variables numèriques (num_critic_for_reviews, duration, actor_3_facebook_likes, actor_2_facebook_likes, actor_1_facebook_likes, gross, num_user_for_reviews, budget, aspect_ratio), aplicarem una *Knn-imputation*, per tal d'aproximar el seu valor. Creiem que és oportú perquè els valors perduts estan a diferents registres i si eliminem qualsevol registre encara que sigui només per un valor perdut es perdrà una part rellevant de la informació. A més a més, hi ha molta informació per imputar de forma que el K-nearest Neighbor funcioni acuradament.

3.2. Identificació i tractament de valors extrems.

Identificació:

Per tal d'identificar aquells valors extrems o outliers els quals semblen no segueixen la mateixa tendència que la resta de dades, hem aplicat dos mesures:

1. Observació de les variables usant un diagrama de caixa o boxplot.
2. Amplitud interquartílica (IQR), sent una mesura de dispersió estadística entre el Q1 i el Q3. Podem identificar els valors extrems usant dos valors llindar: valors menors a Q1 - 1.5 * IQR, i valors majors a Q3 + 1.5 * IQR.



Tractament:

Al observar i revisar els valors per variable o característica del film, podem comprovar com l'obtenció d'aquests és possible. Pot donar-se el cas de que una pel·lícula duri 7 minuts (duració mínima en el dataset) i hagi estat valorada per un alt nombre d'usuaris i tingui una puntuació al IMDB de 7. Per tant el tractament que seguirem serà mantenir aquests valors tal i com es troben en el dataset.

3.3. Transformació i creació de variables

S'ha portat a terme un pas de transformació de variables en el cas de les variables categòriques amb molts valors per tal de poder-les utilitzar de forma més adient. En resum, les variables categòriques queden amb les següents categories:

Variable	Categories
Color	"Color", "Black and white"
Language	"English", "Others"
Genres	"Comedy", "Action", "Drama", "Adventure", "Others"
Continent	"Europe", "America", "Africa", "Asia", "Oceania"
Year	"<2000", ">2000"

4. Anàlisi de les dades.

4.1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

Tal com s'ha comentat a la introducció, cal contemplar les següents etapes:

1. Correlacions de les diferents variables numèriques amb la variable d'interès durada (adjuntant la significació). D'aquesta forma podem detectar associacions fortes i significatives que ens permetin valorar variables amb poder predictiu per introduir-les a l'estudi de regressió lineal.

2. Estudiar la dependència de la durada sobre diferents variables categòriques a partir de contrast d'hipòtesis sobre la diferència de mitjanes de la durada entre els diferents grups de la variable categòrica.

Això implica valorar els supòsits de normalitat i homoscedasticitat per decidir si utilitzar els test paramètrics o no paramètrics.

a) En el cas de ser variables de dues categories s'utilitzaran la *t-student* en cas paramètric i *Mann-Whitney U* en cas no paramètric.

b) En el cas de ser variables de més de dues categories, s'utilitzarà la tècnica *ANOVA* en cas paramètric i *Kruskal Wallis* en cas no paramètric

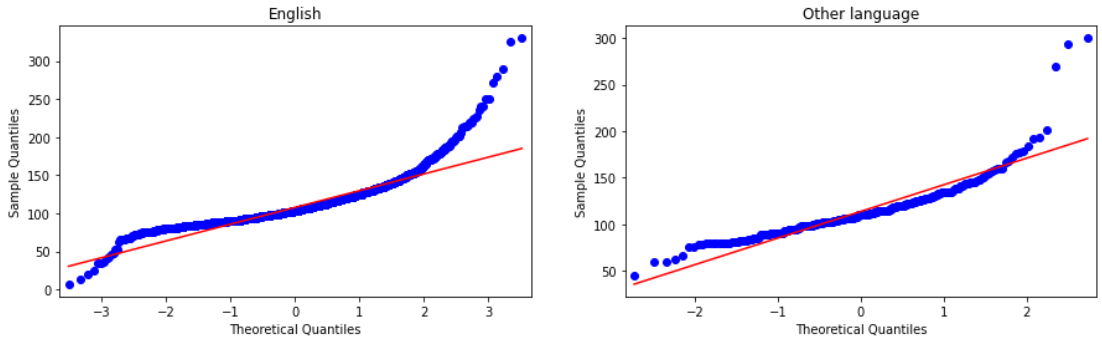
3. Model de regressió lineal múltiple per explicar/predir la durada de les pel·lícules a partir de les característiques amb forta associació trobades als apartats 1 i 2.

4.2. Comprovació de normalitat i homogeneïtat de la variància.

Tal com s'ha especificat en la planificació de l'anàlisi, a l'apartat anterior, en aquest apartat estudiem la normalitat i la homoscedasticitat de la variable durada pels grups que generen les variables categòriques següents: color, genres, language, country i title_year. D'aquesta manera podrem decidir si apliquem test paramètrics (t-test per les dicotòmiques o ANOVA per les de més de dos categories) o test no paramètrics (Mann-Whitney U per les dicotòmiques o Kruskal-Wallis per les de més de dos categories). Finalment, veurem que caldrà utilitzar sempre els test no paramètrics.

***Variable idioma**:* Variable transformada en dos grups ("English", "No English").

Normalitat: La variable durada no és normal ni al grup d'idioma anglès ni al grup d'idioma no anglès.



Tal com indica la següent taula, els valors P del test de normalitat propers a zero confirmen la no normalitat de les dades.

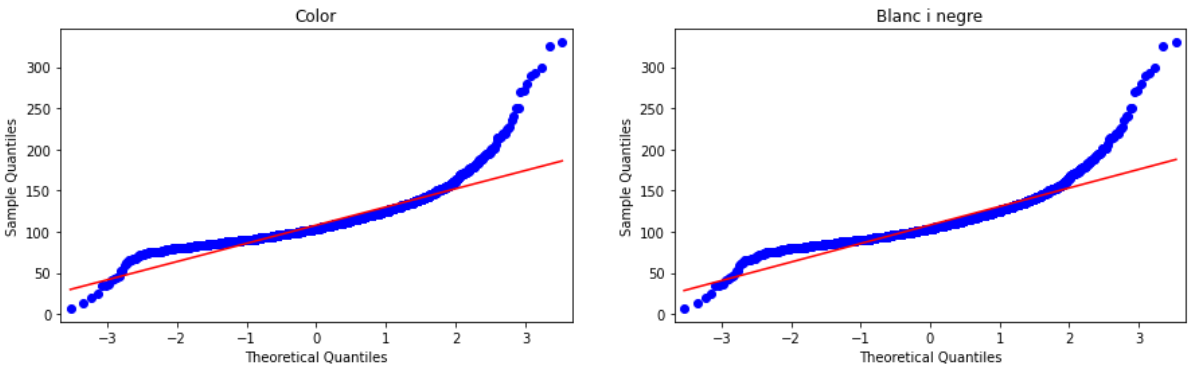
Grups	Test Normalitat (D'Agostino and Pearson's)	
	Valor observat	Valor P
Language="English"	2270,97	<0,001
Language="Others"	201,66	<0,001

Homoscedasticitat: Tal com indica la taula, el test de Levene per igualtat de variàncies indica una probabilitat d'equivocar-nos molt baixa si refusem i per tant podem concloure que les variàncies són diferents.

Test (Homoscedasticitat)	Valor observat	Valor P
Levene	13,24	<0,001

Per tant, caldrà utilitzar el test no paramètric de *Mann-Whitney U*

***Variable color**:* Variable transformada en dos grups ("Color", "Black and white")
Normalitat: La variable durada no és normal en cap grup.



Tal com indica la següent taula, els valors P del test de normalitat propers a zero confirmen la no normalitat de les dades.

Grups	Test Normalitat (D'Agostino and Pearson's)	
	Valor observat	Valor P
Color="Color"	2514,67	<0,001
Color="Black and white"	2528,58	<0,001

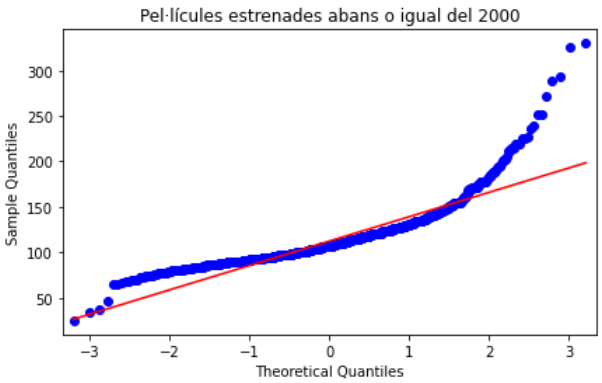
Homoscedasticitat: Tal com indica la taula, el test de Levene per igualtat de variàncies indica una probabilitat d'equivocar-nos molt alta si refusem i per tant podem concloure que les variàncies són iguals a nivell poblacional.

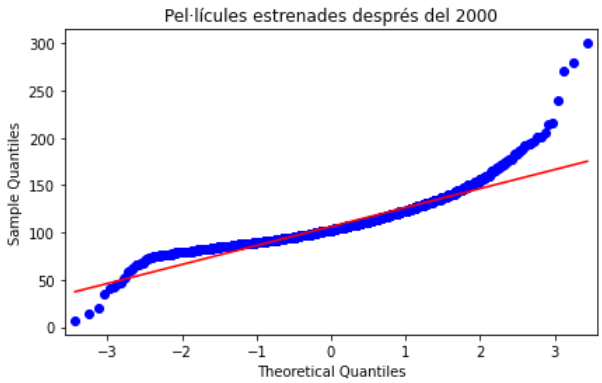
Test (Homoscedasticitat)	Valor observat	Valor P
Levene	0,7105	0,399

Com no es compleix la normalitat caldrà utilitzar el test no paramètric de *Mann-Whitney U*

Variable year: Variable transformada en dos grups ("<=2000", ">2000")

Normalitat: La variable durada no és normal en cap dels 2 grups.





Tal com indica la següent taula, els valors P del test de normalitat propers a zero confirmen la no normalitat de les dades.

Grups	Test Normalitat (D'Agostino and Pearson's)	
	Valor observat	Valor P
Year="<=2000"	841,16	<0,001
Year=">2000"	1368,81	<0,001

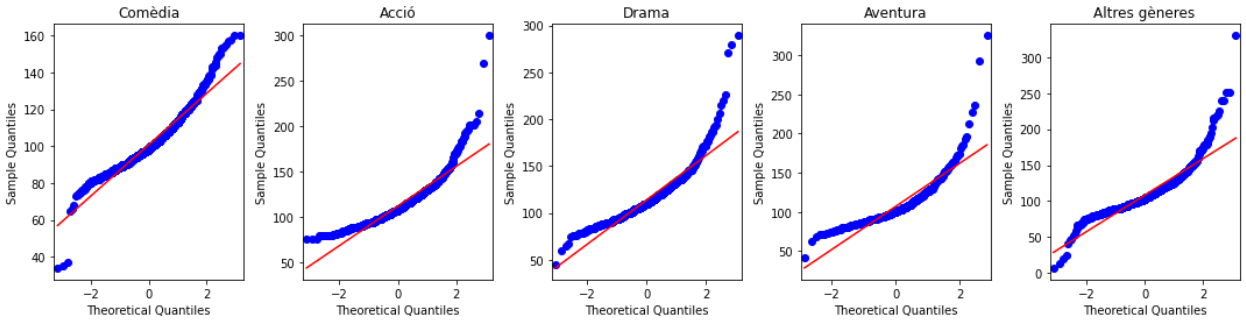
Homoscedasticitat: Tal com indica la taula, el test de Levene per igualtat de variàncies indica una probabilitat d'equivocar-nos molt baixa si refusem i per tant podem concloure que les variàncies són diferents.

Test (Homoscedasticitat)	Valor observat	Valor P
Levene	2209487	<0,001

Per tant, caldrà utilitzar el test no paramètric de *Mann-Whitney U*

Variable gènere: Variable transformada en cinc grups ("Comedy", "Action", "Drama", "Adventure", "Other genre") .

Normalitat: La variable durada no és normal en cap dels 5 grups.



Tal com indica la següent taula, els valors P del test de normalitat propers a zero confirmen la no normalitat de les dades.

Grups	Test Normalitat (D'Agostino and Pearson's)	
	Valor observat	Valor P
Gènere="Comedy"	180,28	<0,001
Gènere="Action"	539,40	<0,001
Gènere="Drama"	444,75	<0,001
Gènere="Adventure"	312,06	<0,001
Gènere="Other-gen"	520,33	<0,001

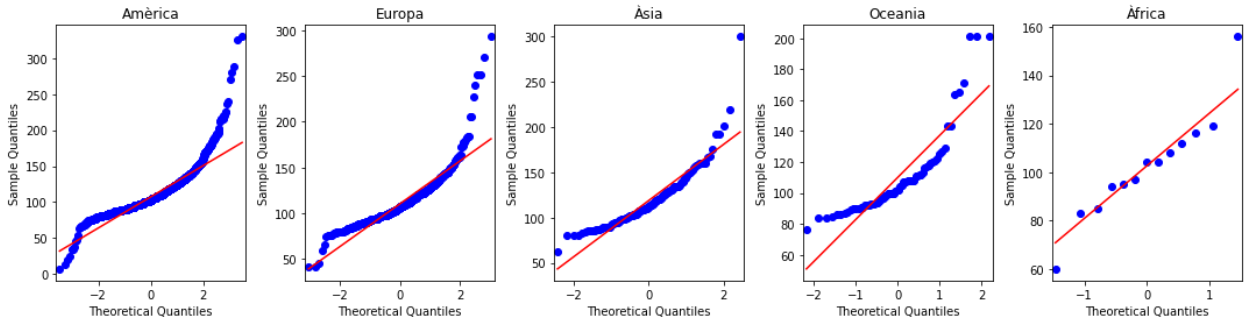
Homoscedasticitat: Tal com indica la taula, el test de Levene per igualtat de variàncies indica una probabilitat d'equivocar-nos molt baixa si refusem i per tant podem concloure que les variàncies són diferents.

Test (Homoscedasticitat)	Valor observat	Valor P
Levene	37,48	<0,001

Per tant, caldrà utilitzar el test no paramètric de *Kruskal-Wallis*

Variable continent:

Normalitat: La variable durada no és normal en cap dels 5 grups.



Tal com indica la següent taula, els valors P del test de normalitat propers a zero confirmen la no normalitat de les dades excepte pel cas del continent africà que la duració seria normal.

Grups	Test Normalitat (D'Agostino and Pearson's)	
	Valor observat	Valor P
Continent="America"	1951,28	<0,001
Continent="Europe"	464,32	<0,001
Continent="Asia"	80,23	<0,001
Continent="Oceania"	38,28	<0,001
Continent="Africa"	3,97	0,137

Homoscedasticitat: Tal com indica la taula, el test de Levene per igualtat de variàncies indica una probabilitat d'equivocar-nos molt baixa si refusem i per tant podem concloure que les variàncies són diferents.

Test (Homoscedasticitat)	Valor observat	Valor P
Levene	5,89	<0,001

Per tant, caldrà utilitzar el test no paramètric de *Kruskal-Wallis*

4.3. Aplicació de proves estadístiques per comparar grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

1. Proves de significació (contrast d'hipòtesis):

Variable idioma: Tal com hem vist en la part del supòsits de normalitat i homoscedasticitat, cal aplicar un test ***U de Mann-Whitney.***

Test	Valor observat	Valor P
U de Mann-Whitney	619520,5	<0,001

Tal com es pot observar a la taula anterior, obtenim un valor observat de 619520,5 amb un valor P associat molt petit, inferior al 0,001. Per tant, podem afirmar que la distribució de la variable durada en els films dels diferents grups no són iguals. És a dir, la variable en qüestió afecta significativament a la durada del film i per tant, serà inclosa a la regressió.

Variable color**:* Tal com hem vist en la part del supòsits de normalitat i homoscedasticitat, cal aplicar un test ***U de Mann-Whitney.

Test	Valor observat	Valor P
U de Mann-Whitney	11526756	0,375

Tal com es pot observar a la taula anterior, obtenim un valor observat de 11526756 amb un valor P associat alt. Per tant, no podem afirmar que la distribució de la variable durada en els films dels diferents grups no són iguals. És a dir, la variable en qüestió NO afecta a la durada del film i per tant, no la inclourem en la regressió.

Variable year**:* Tal com hem vist en la part del supòsits de normalitat i homoscedasticitat, cal aplicar un test ***U de Mann-Whitney.

Test	Valor observat	Valor P
U de Mann-Whitney	2209487	<0,001

Tal com es pot observar a la taula anterior, obtenim un valor observat de 2209487 amb un valor P associat molt petit, inferior al 0,001. Per tant, podem afirmar que la distribució de

la variable durada en els films dels diferents grups no són iguals. És a dir, la variable en qüestió afecta significativament a la durada del film i per tant, serà inclosa a la regressió.

Variable gènere**:* Tal com hem vist en la part del supòsits de normalitat i homoscedasticitat, cal aplicar un test ***Kruskal-Wallis.

Test	Valor observat	Valor P
Kruskal-Wallis	302,43	<0,001

Tal com es pot observar a la taula anterior, obtenim un valor observat 302,43 amb un valor P associat molt petit, inferior al 0,001. Per tant, podem afirmar que la distribució de la variable durada en els films dels diferents grups no són iguals. És a dir, la variable en qüestió afecta significativament a la durada del film i per tant, serà inclosa a la regressió.

Variable continent**:* Tal com hem vist en la part del supòsits de normalitat i homoscedasticitat, cal aplicar un test ***Kruskal-Wallis.

Test	Valor observat	Valor P
Kruskal-Wallis	36,21	<0,001

Tal com es pot observar a la taula anterior, obtenim un valor observat de 36,21 amb un valor P associat molt petit, inferior al 0,001. Per tant, podem afirmar que la distribució de la variable durada en els films dels diferents grups no són iguals. És a dir, la variable en qüestió afecta significativament a la durada del film i per tant, serà inclosa a la regressió.

2. Correlacions:

Primer hem calculat la correlació de Pearson i el p-value associat usant la funció *personr* entre la variable *duration* i la resta de variables numèriques. Les 10 variables que mostren una correlació significativa són:

	correlation	pvalue
num_user_for_reviews	0.352534	9.677374e-144
num_voted_users	0.342481	2.974620e-135
imdb_score	0.342185	5.229846e-135
num_critic_for_reviews	0.250943	1.965519e-71
gross	0.232756	1.952738e-61
movie_facebook_likes	0.218328	4.280753e-54
director_facebook_likes	0.173426	1.770237e-34
actor_2_facebook_likes	0.137036	4.996309e-22
actor_3_facebook_likes	0.132285	1.268529e-20
cast_total_facebook_likes	0.126943	4.196485e-19

Per tant, inclourem aquestes variables a la regressió encara que caldria controlar la multicol·linealitat per decidir quines variables acaben formant part del model.

3. Model de regressió lineal múltiple:

Seguint la planificació inicial, és de gran interès poder predir la duració de les pel·lícules segons les seves característiques. Per tant calcularem un model de regressió lineal múltiple utilitzant regressors quantitatius i qualitatius per tal de realitzar prediccions sobre la duració dels films.

Categories de referència per les variables categòriques:

Language: ‘Other language’, Gènere: ‘Other-gen', Year: '<=2000', Continent: ‘Europe’.

Els resultats de la regressió és el següent:

OLS Regression Results

Dep. Variable:	duration	**R-squared (uncentered):**	0.964
Model:	OLS	**Adj. R-squared (uncentered):**	0.963
Method:	Least Squares	**F-statistic:**	5627.
Date:	Sun, 03 Jan 2021	**Prob (F-statistic):**	0.00
Time:	19:48:29	**Log-Likelihood:**	-21949.
No. Observations:	4913	**AIC:**	4.394e+04
Df Residuals:	4890	**BIC:**	4.409e+04
Df Model:	23		
Covariance Type:	nonrobust		

I, respecte els coeficients:

	coef	std err	t	P>|t|	[0.025	0.975]
num_critic_for_reviews	-0.0130	0.004	-2.981	0.003	-0.022	-0.004
director_facebook_likes	0.0005	0.000	4.347	0.000	0.000	0.001
actor_3_facebook_likes	-0.0018	0.000	-3.703	0.000	-0.003	-0.001
actor_2_facebook_likes	-0.0015	0.000	-4.964	0.000	-0.002	-0.001
actor_1_facebook_likes	-0.0016	0.000	-5.538	0.000	-0.002	-0.001
gross	1.982e-08	6.61e-09	2.997	0.003	6.85e-09	3.28e-08
num_voted_users	-2.961e-05	4.32e-06	-6.853	0.000	-3.81e-05	-2.11e-05
cast_total_facebook_likes	0.0016	0.000	5.576	0.000	0.001	0.002
num_user_for_reviews	0.0163	0.001	11.571	0.000	0.014	0.019
budget	4.076e-09	1.56e-09	2.617	0.009	1.02e-09	7.13e-09
imdb_score	11.1567	0.200	55.664	0.000	10.764	11.550
aspect_ratio	4.5023	0.382	11.798	0.000	3.754	5.250
movie_facebook_likes	5.788e-05	2.26e-05	2.566	0.010	1.37e-05	0.000
English	18.1815	1.287	14.125	0.000	15.658	20.705
Action	8.3247	0.904	9.211	0.000	6.553	10.097
Adventure	2.5724	1.196	2.151	0.032	0.227	4.917
Comedy	1.6322	0.860	1.898	0.058	-0.054	3.318
Drama	8.4080	0.940	8.944	0.000	6.565	10.251
>2000	-0.0911	0.709	-0.128	0.898	-1.481	1.299
Africa	-3.2851	5.926	-0.554	0.579	-14.903	8.333
America	1.2892	0.857	1.505	0.132	-0.390	2.968
Asia	22.4868	2.105	10.681	0.000	18.360	26.614
Oceania	-0.9100	2.686	-0.339	0.735	-6.177	4.357

Omnibus:	2042.000	**Durbin-Watson:**	1.830
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	29901.241
Skew:	1.585	**Prob(JB):**	0.00
Kurtosis:	14.663	**Cond. No.**	3.99e+09

Com podem observar, el model de regressió obtingut té un coeficient de determinació (R^2) alt (0.964), per tant podem determinar que aquest serà un bon model per realitzar prediccions.

5. Representació dels resultats a partir de taules i gràfiques.

Tant en aquest document de respostes com en el notebook de Jupyter, podem observar taules i gràfiques obtingudes al llarg de la pràctica per tal de poder observar i extreure conclusions més ràpidament.

6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre el problema?

En aquesta pràctica hem aplicat tres probes estadístiques sobre el dataset d'interès el qual contenia informació sobre diferents pel·lícules, prèviament tractat per seleccionar aquelles variables informatives per aconseguir l'objectiu inicial. Per cada variable hem estudiat la seva normalitat i la homogeneïtat de la variància.

Després, a partir de l'anàlisi de correlació i el contrast d'hipòtesis ens ha permès identificar quines d'aquestes variables tenen un pes més significatiu a l'hora de determinar la duració de les pel·lícules. Finalment, el model de regressió lineal que hem obtingut pot ser de gran utilitat per tal de realitzar prediccions sobre la duració de la pel·lícula a partir d'unes característiques determinades. Per tant, podem concloure que amb els resultats que hem obtingut podem respondre el problema.

7. Codi: cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

El codi s'entrega en Python a través d'un notebook de Jupyter. D'aquesta manera es pot fàcilment seguir tot el procediment que hem seguit per realitzar la pràctica.

Contribucions	Firma
Investigació prèvia	Silvia Galan, David Roche
Redacció de les respostes	Silvia Galan, David Roche
Desenvolupament del codi	Silvia Galan, David Roche