SECONDO HOMEWORK STATISTICA COMPUTAZIONALE

'E lei si sente più Spiderman o Batman?'

MARVEL vs DC

ADEZIO Giuditta, FOLLADOR Francesca

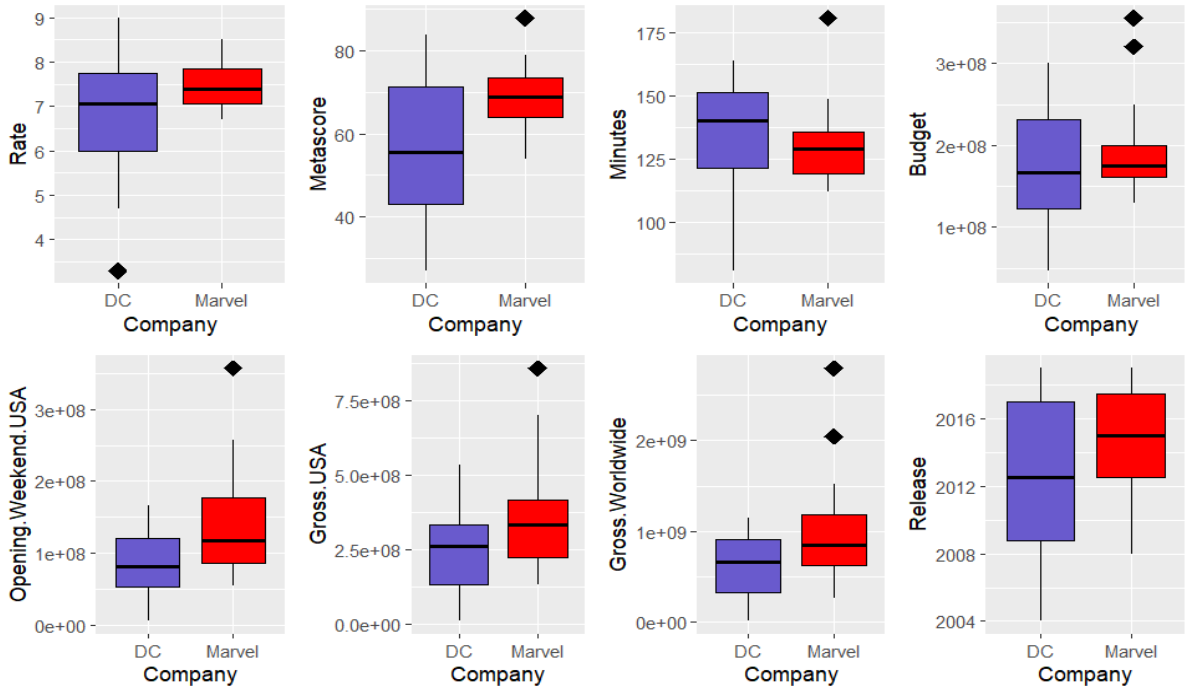
"I film sono il modo per vedere il mondo attraverso gli occhi di qualcun altro."

Il dataset a nostra disposizione (reperibile su Kaggle) contiene 39 unità statistiche riguardanti la battaglia che si protrae ormai da anni tra i film della Marvel e quelli della DC, usciti dal 2000 al 2019. In particolare contiene le seguenti 10 variabili (al netto della prima colonna che è un intero da 1 a 39):

- Original Title (stringa, titolo originale del film)
- Company (stringa, Marvel o DC)
- Rate (decimale, voto assegnato al film da IMDB)
- Metascore (intero, voto assegnato al film da metacritic.com)
- Minutes (stringa, durata del film)
- Release (intero, anno in cui il film è stato pubblicato)
- Budget (stringa, budget in dollari utilizzato per realizzare il film)
- Opening.weekend.USA (intero, incasso lordo del film nel primo weekend dalla pubblicazione negli Stati Uniti)
- Gross.USA (intero, incasso totale del film in USA)
- Gross.Worldwide (decimale, incasso totale del film nel mondo)

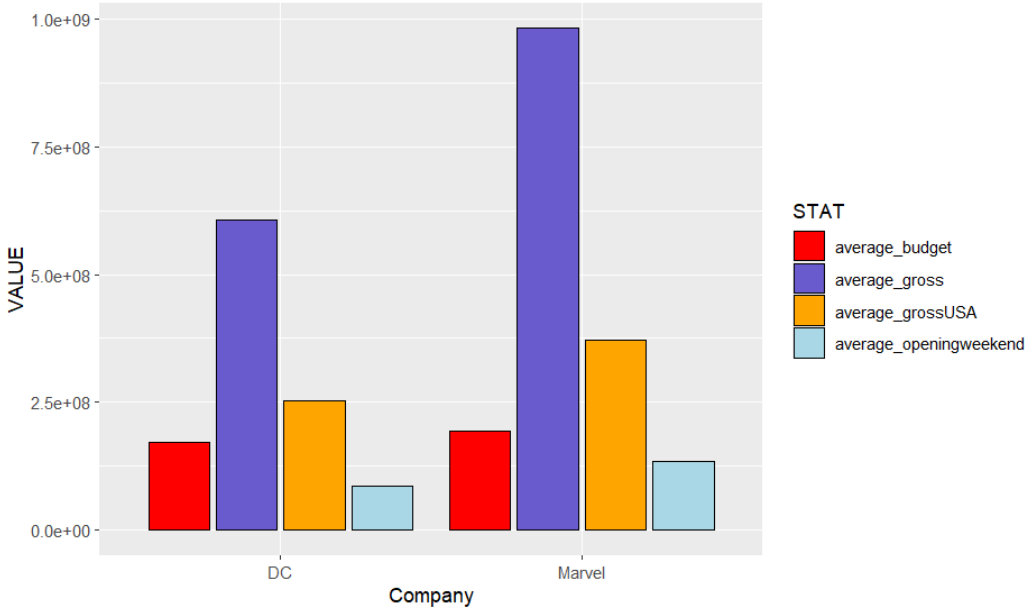
ANALISI ESPLORATIVA DEI DATI

Per la nostra analisi abbiamo reso il dataset tidy e abbiamo valutato la presenza di valori mancanti e outliers. Per quanto riguarda i primi nel dataset preso in considerazione non ce ne sono, invece per analizzare i secondi come prima cosa abbiamo realizzato dei boxplot.

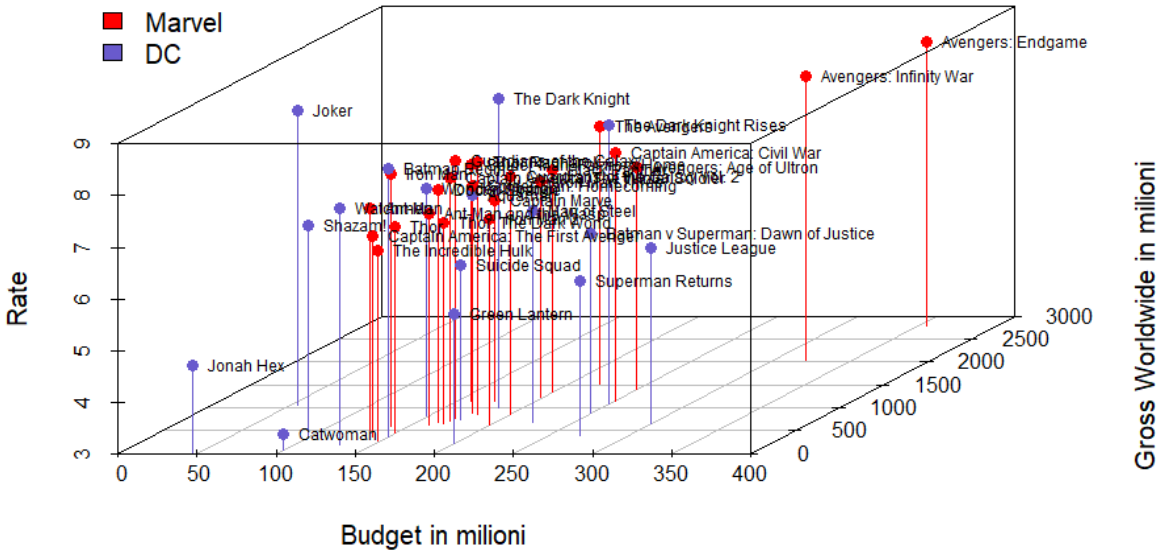


Avendo notato la presenza di alcuni outliers, abbiamo eseguito un'analisi più approfondita notando che il film Avengers: Endgame rappresenta un outlier per budget, incasso totale nel mondo e nel weekend di uscita del film negli USA molto più elevati rispetto a quelli dei restanti film. Altri punti critici sono rappresentati dai film Catwoman e Jonah Hex rispettivamente con un Rate e una durata del film molto bassi.
Da un'analisi generale si può osservare che i film della Marvel hanno in media punteggi, budget e incassi più elevati di quelli della DC ma durata dei film in media inferiore. La variabilità delle variabili dei film della DC è maggiore rispetto a quella dei film della Marvel.

Di seguito un grafico riassuntivo delle statistiche coerente con quanto detto in precedenza.

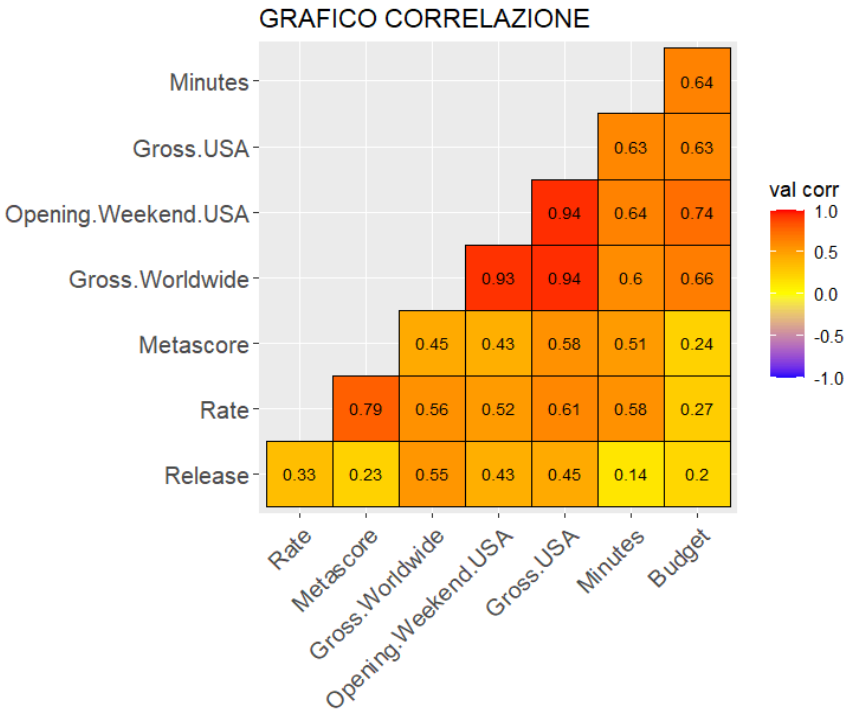


Prese delle variabili ai nostri occhi di particolare interesse, abbiamo deciso di mettere in relazione Rate, Budget e Gross Worldwide con uno scatterplot 3D. Dopo il film "Jonah Hex" (secondo più basso per Rate) notiamo che il secondo film per budget più basso è "Catwoman", entrambi film della DC, come lo è anche "The Dark Knight" che però riporta il Rate più elevato (seguito da "Joker"). Su R è anche presente lo scatterplot 3D interattivo.



ANALISI DELLE CORRELAZIONI

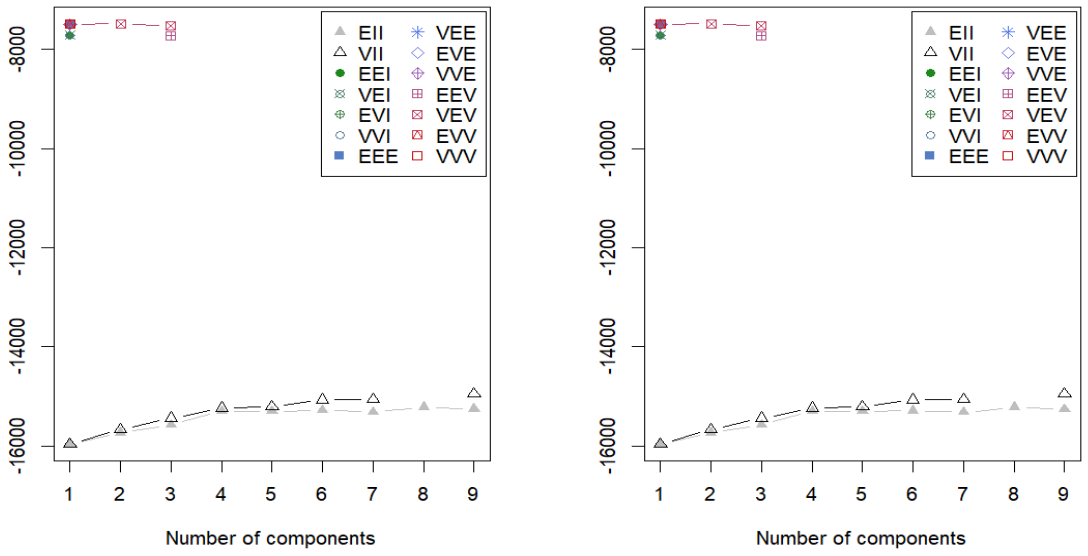
Da una prima analisi possiamo vedere che le variabili che in assoluto risultano più correlate sono Opening weekend USA e Gross USA, che è anche ugualmente correlata con Gross worldwide (correlata naturalmente con opening weekend USA).
Altre variabili abbastanza correlate sono Rate e Metascore, come potevasi immaginare.
Le meno correlate risultano essere Minutes e Release.
Tutte le variabili sono correlate positivamente.



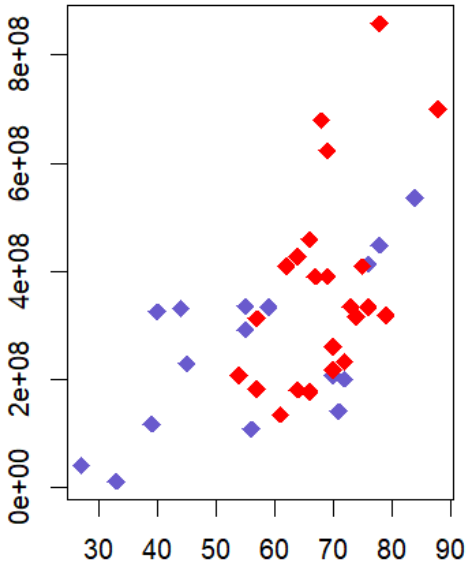
A seguito di quanto evinto dal grafico sulle correlazioni, abbiamo deciso di procedere con un'analisi delle componenti principali per vedere se fosse necessario nelle indagini a seguire considerare tutte le variabili. Il risultato è il seguente: le prime tre componenti principali spiegano più dell'80% della varianza. L'analisi potrebbe essere fatta anche solo con le variabili Metascore, Gross USA e Release ma per completezza la effettueremo prima con tutte le variabili.

MODEL BASED CLUSTERING

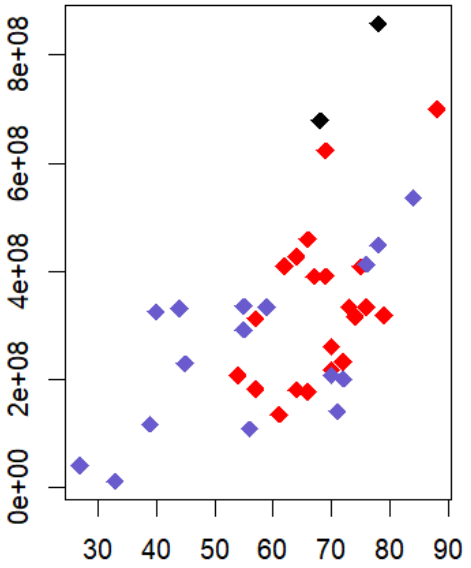
Adesso siamo passate alla clusterizzazione dei dati con un model based clustering.
Riportiamo un grafico per visualizzare il modello che è stato scelto tramite BIC e ICL.



Da entrambi i criteri valutati è emerso che il modello più indicato per i nostri dati è un VEV (ellipsoidal with equal shape) con due componenti con ICL = 7484.416 e BIC = 7484.402.
I due cluster sono rispettivamente composti da 21 e 18 unità statistiche.
Conoscendo le vere etichette (23 Marvel, 16 DC), abbiamo eseguito un confronto con quelle ottenute dalla clusterizzazione dal quale è emerso che il CER (error rate) è pari a 0.05128205 (quindi l'accuracy è pari a 0.948718) e l'ARI (adjusted rate index) è pari a 0.8002972.
In conclusione da questi dati e da ulteriori analisi (confusion matrix) possiamo dire che il modello di clusterizzazione trovato è un buon modello.



1. TRUE CLASSIFICATION

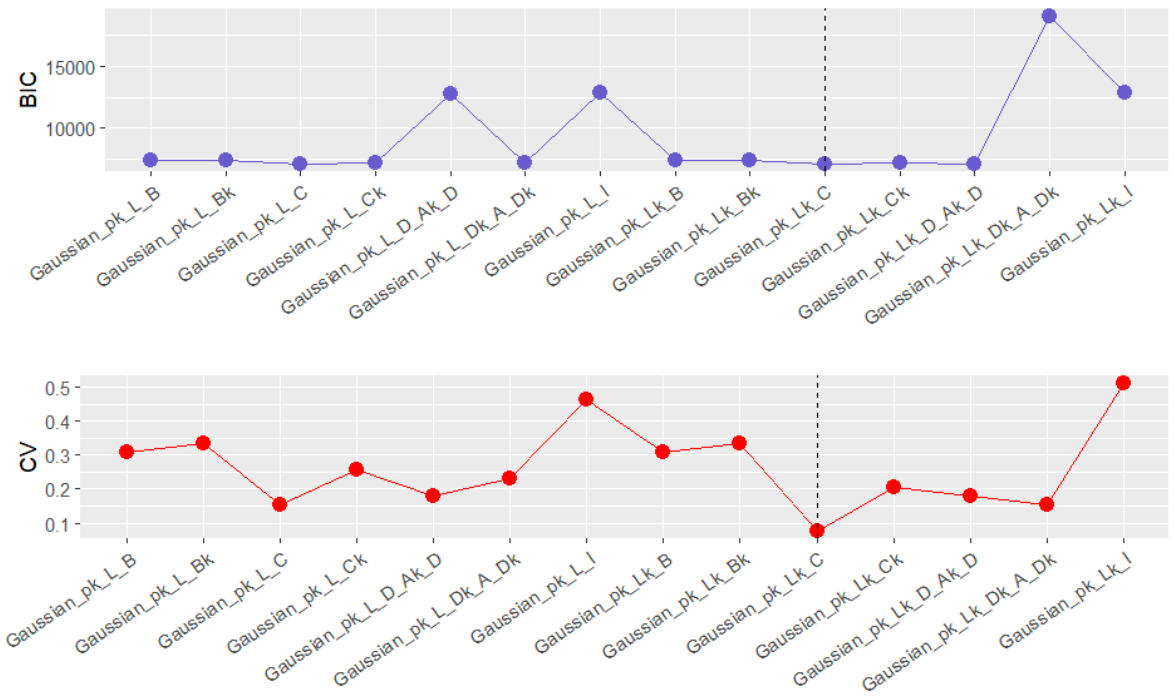


2. MODEL BASED CLUSTERING

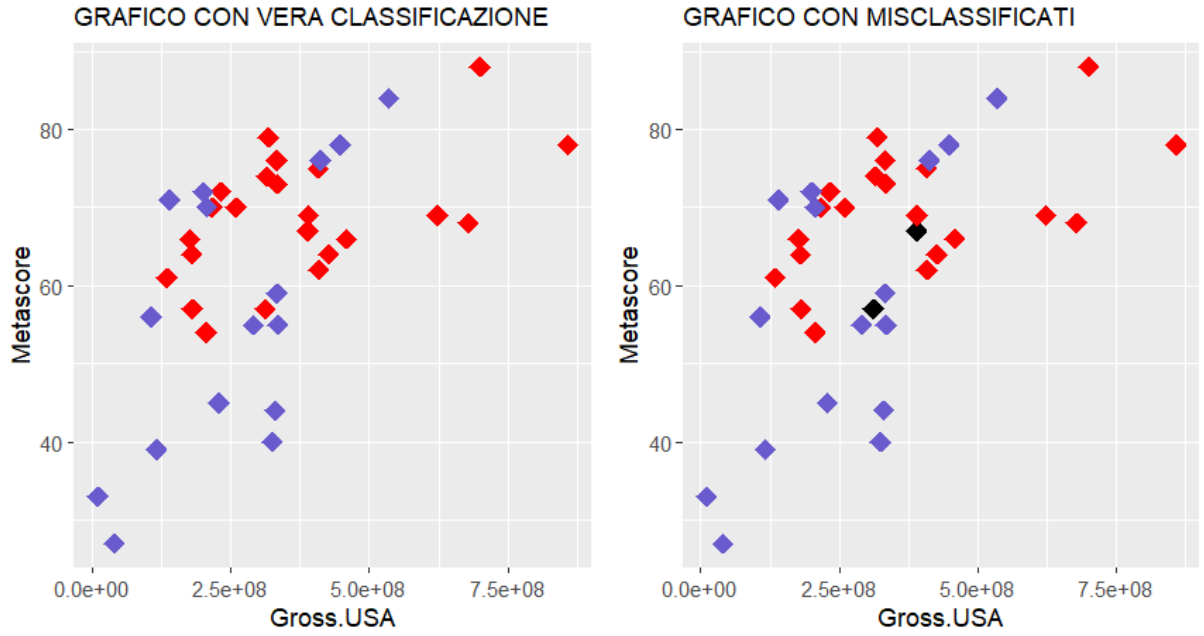
Guardando il grafico di confronto tra le vere etichette e quelle calcolate dal model based clustering vediamo che le unità misclassificate sono due 'Avengers: Endgame' e 'Avengers: Infinity War'.

A questo punto abbiamo eseguito la clusterizzazione con le variabili trovate con l'analisi delle componenti principali per effettuare una comparazione. Il modello risultante è lo stesso (con BIC = 2143.543 e ICL = 2147.554) ma la clusterizzazione risulta peggiore poiché il CER è pari a 0.3846154 e l'ARI è pari a 0.02797718. Le unità sono state classificate in questo modo: 25 nel primo gruppo e 14 nel secondo gruppo.

MODEL BASED CLASSIFICATION CON EDDA
Avendo a disposizione le vere etichette, abbiamo classificato il modello utilizzando la famiglia dei classificatori EDDA (Eigenvalue Decomposition Discriminant Analysis).
Il modello scelto, eseguendo la classificazione più volte, è un Gaussian_pk_Lk_C cioè un VEE (free proportion, free volume, equal shape e equal orientation) con due cluster e con CV = 0.1025641 e BIC = 7126.291.
I successivi due modelli più frequenti (a quello che è risultato il migliore) sempre con 2 cluster, risultano essere Gaussian_pk_Lk_D_Ak_D (VVE) e Gaussian_pk_L_C (EEE, LDA).
Nel grafico sottostante un'analisi grafica con verifica del modello scelto, in cui si nota che per entrambi i criteri (BIC e CV) il modello scelto è lo stesso, cioè un Gaussian_pk_Lk_C.



Dopo aver allenato il classificatore al netto di un campione casuale di 7 unità statistiche, abbiamo testato il modello risultante (Gaussian_pk_Lk_C) usando le unità statistiche escluse in precedenza.
Confrontando il vero gruppo di appartenenza delle unità statistiche considerate con quello stimato dalla classificazione si può concludere che quello trovato sia un buon modello.
Confronto grafico del modello trovato con EDDA con le vere etichette:

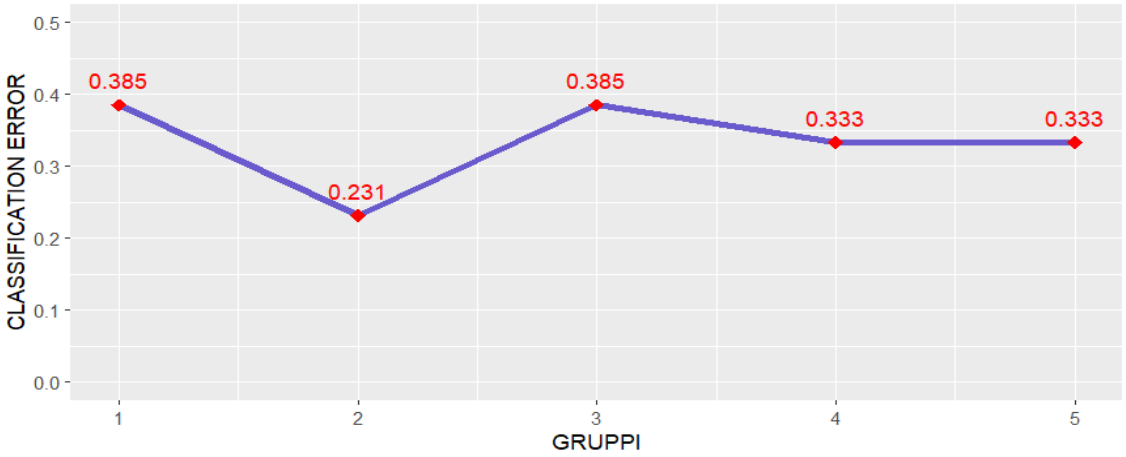


MODEL BASED CLASSIFICATION CON MDA

Per completezza, nonostante non fosse necessario dato che con l'EDDA è stato trovato un buon modello con un class error pari a circa 0.05, abbiamo ora eseguito una Mixture Discriminant Analysis, allenando il modello al netto del test set estratto in precedenza e poi

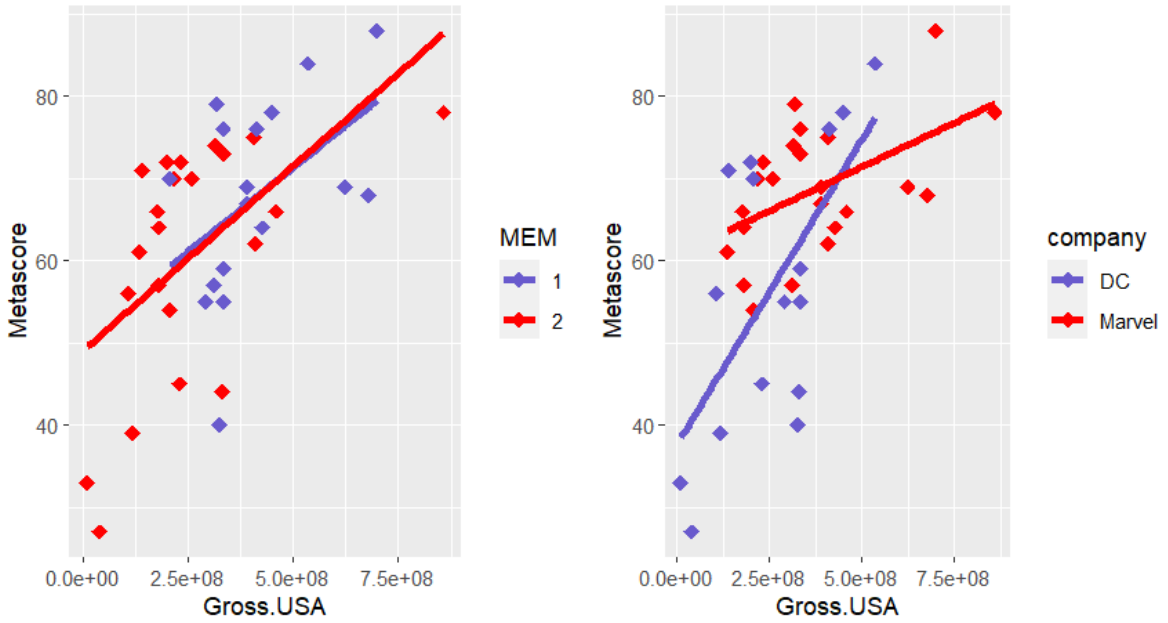
testandolo con quelle stesse unità statistiche. Il risultato di questa analisi, scegliendo a seconda del BIC, conferma il modello trovato in precedenza con la clusterizzazione (VEV) per entrambe le misture interne, ma con due cluster di, rispettivamente, 14 e 18 unità statistiche. Possiamo concludere che non era necessario calcolare il modello con MDA.

In seguito abbiamo implementato, imponendo il modello trovato sia con MDA che con clusterizzazione (VEV), una Cross Validation V Fold che conferma la scelta di due gruppi.



FINITE MIXTURES OF REGRESSION MODEL

Infine, anche utilizzando il finite mixture of regression model, vediamo che il modello scelto è sempre composto da due cluster. Il class error dato dal regression model però risulta essere molto alto (0.4358974, unità misclassificate 17) quindi il modello stimato non si adatta bene ai nostri dati, come si evince anche dal grafico sottostante.



Concludiamo dicendo che tutti i modelli utilizzati dividono il dataset a nostra disposizione in due gruppi e sembrano adattarsi bene, al netto della Mixture of Expert Models, ai nostri dati.

ADEZIO Giuditta, FOLLADOR Francesca
Anno Accademico 2023/2024, SSE