Use no binary dataset with all features:

features : ["production\_companies\_index", "production\_countries\_index", "genres\_index", "release\_date", "budget", "avg\_rating", "runtime", "cast\_index", "director\_index"]

Validation score: 76358691.97083652

(9,[0,1,2,3,4,5,6,7,8],[0.09808484807668617,0.07110851225925555,0.06629424918717544,0.029178761966011085,0.5855712676145417,0.022583241286527618,0.09757347512978025,0.01805234269588688,0.011553301784135207])

The most important is budget;

So delete that and run again

features: ["production\_companies\_index", "production\_countries\_index", "genres\_index", "release\_date", "avg\_rating", "runtime", "cast\_index", "director\_index"]

Validation score:65131717.90827548

(8,[0,1,2,3,4,5,6,7],[0.16904821856010835,0.13614727936584062,0.25198634741202214,0.07695948737150389,0.07531185035493203,0.20534151872116438,0.05577687180502463,0.029428426409403935])

The most important one is genre;

okay delete that and run again

feature:["production\_companies\_index", "production\_countries\_index", "release\_date",

"avg\_rating", "runtime", "cast\_index", "director\_index"]

Validation score:65461267.910716

(7,[0,1,2,3,4,5,6],[0.19459385892911715,0.22697098593733211,0.08128792536908191,0.07961636700296645,0.23495332515925918,0.12302562338179499,0.05955191422044821])

The most important one is runtime;

okay delete that and run the final time

feature:"production\_companies\_index", "production\_countries\_index", "release\_date",

"avg\_rating", "cast\_index", "director\_index"

Validation score:74945513.48235223

(6,[0,1,2,3,4,5],[0.2598357226311524,0.15457885240041994,0.16884199258823046,0.14342913808872537,0.15958053662968022,0.11373375766179143])

for this time is production\_company,

And now we get all for no binary data, that is : budget; genres; runtime; production\_company.

Now we use binary data to try this :

First:

Feature:"key\_company", "key\_country", "genres\_index", "release\_date", "budget", "avg\_rating", "runtime", "key\_cast", "key\_director"

Validation score:69497556.9664307

(9,[0,1,2,3,4,5,6,7,8],[0.011763529641655815,0.006355344201469879,0.09168820740853638,0.06708964184577174,0.6831693094891046,0.02588532022720615,0.10146169437697024,0.008703694142560461,0.003883258666724847])

The most important one is budget,

Second time:

Feature:"key\_company", "key\_country", "genres\_index", "release\_date", "avg\_rating", "runtime", "key\_cast", "key\_director"

Validation score: 64931143.35337234

(8,[0,1,2,3,4,5,6,7],[0.023236811009175354,0.039254247174437956,0.3256739430749291,0.1289384669451495,0.133590862945104,0.3219851097714746,0.018076833945932454,0.009243725133797013])

The most important one is genres,

Third time:

Feature:"key\_company", "key\_country", "release\_date", "avg\_rating", "runtime", "key\_cast", "key\_director"

Validation score:64628196.22508063

(7,[0,1,2,3,4,5,6],[0.07969739040971292,0.056832876362252496,0.24669147527961802,0.1688777428529042,0.40534147625683203,0.03458359987996743,0.007975438958712942])

The most important one is runtime,

Final time:

Feature:"key\_company", "key\_country", "release\_date", "avg\_rating", "key\_cast", "key\_director"

Validation score: 75294599.04433846

(6,[0,1,2,3,4,5],[0.09290157381040116,0.07384406566107993,0.46239973281628444,0.30689377795147144,0.04735857988249041,0.016602269878272645])

The most important one is release\_date,

Now we combine two dataset and get the same output:

Most important features are: budget; genres; and runtime

Now we check the correlation\_coefficient :

For no binary data:

profit-company0.008800985496141906

profit-country-0.04732083426375101

profit-director0.01324332914921379

profit-genres0.10028918630312438

profit-cast0.029877966234940205

profit-runtime0.07482394913857125

profit-avg\_rating0.0016589988844411376

profit-budget0.549263170953318

profit-release\_date0.01702159350792003

Top three: budget; genres, runtime

For binary data:

profit-company0.03954431472445275

profit-country0.0024379757508900183

profit-director0.003971610399190495

profit-genres0.10028918630312438

profit-cast0.032963255981083336

profit-runtime0.07482394913857125

profit-avg\_rating0.0016589988844411376

profit-budget0.549263170953318

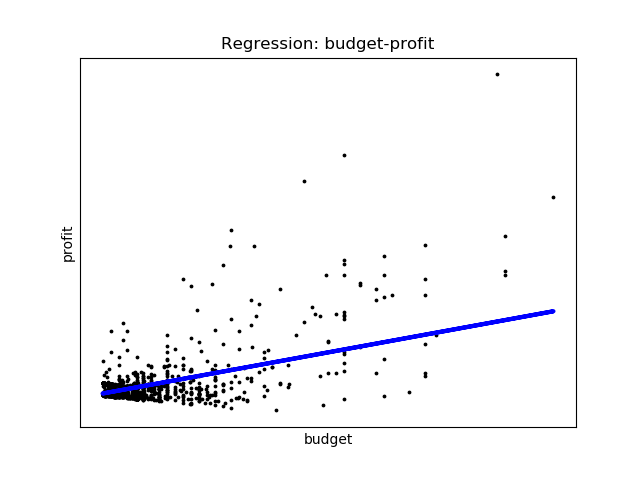
profit-release\_date0.01702159350792003

Top three: budget; genres, runtime

Two methods, two datasets get the same result: budget; genres, runtime

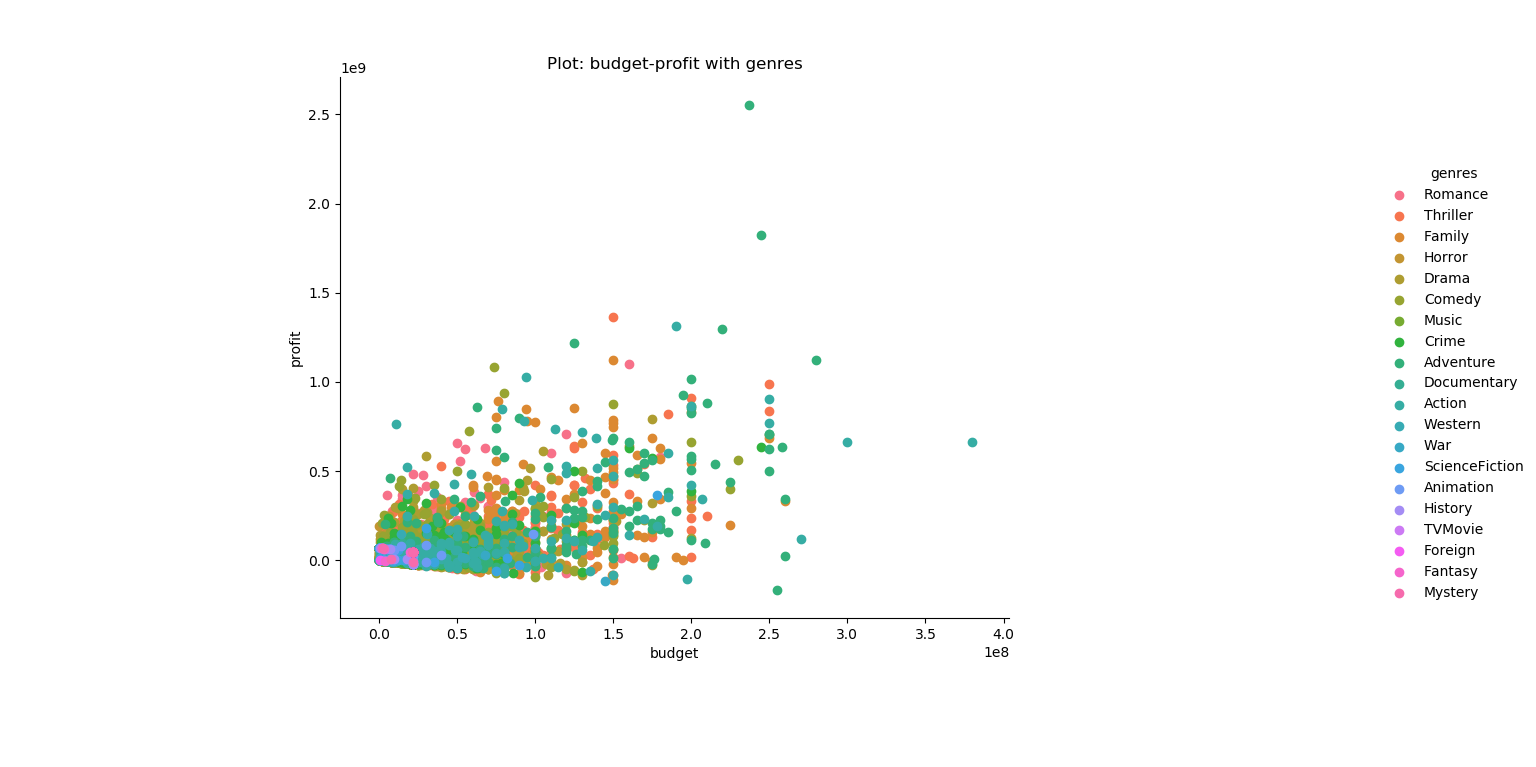
Visualization:

Profit – Budget:



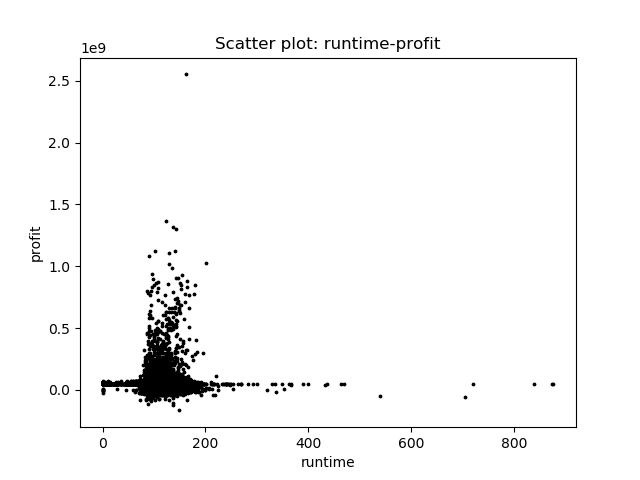
Usually most profitable movies are with high budget;

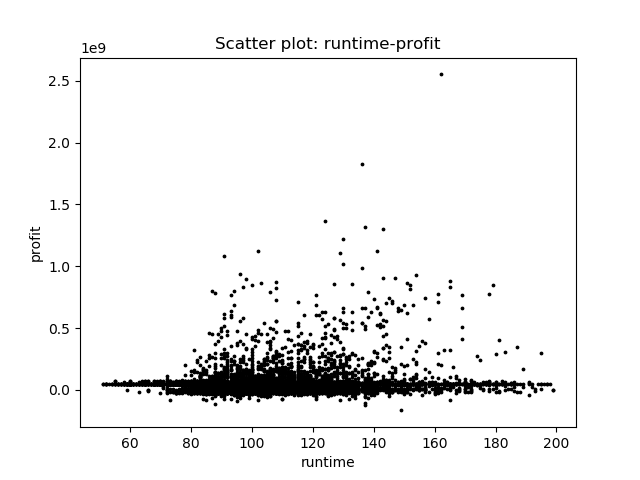
Profit – Genres



Usually, profitable movies are crime movies, thriller movie, and adventure movies;

Profit – Runtime





Usually, high profitable movies are with the runtime between 90-180 minutes.