Name: Adil A Kumar

Social Network Analytics – Empirical Assignment 4

MSBA 2020

1. First, we want to know if filmmakers that engage in collaborations with one another are more innovative or not. We can measure innovation through the number of new, never-before-seen keywords that are used in a film. We can also measure innovation through the number of new combinations of existing keywords that are used in a film. To account for the natural time cycle of the production process, consider a keyword or combination to be “new” if it has been introduced within the last three years.

We also want to know what kinds of collaborations contribute to innovation: are collaborations between large, “generalist” production companies more innovative? Or, are collaborations between large producers and more specialized, smaller producers more innovative?

For this first question, consider two different measures of identifying whether a firm is a generalist or not. Base one measure on the scale of a company’s productions: consider a production company to be a generalist if it is in the top quartile of the number of films released by producers that year. In general, a producer will be classified as a generalist if it makes more than one film in a year.

Base the second measure on a company’s global coreness in the collaboration network: consider a production company to be a generalist if it is in the top quartile of coreness (eigenvector centrality) over its last ten observations. If a company has fewer than ten observations, treat the non-existent observations as zeros.

(A) Classify each film by the type of collaboration that it represents. There should be five

types for each measure of generalism:

i. Peripheral solo productions: films made by a single specialist

ii. Central solo productions: films made by a single generalist

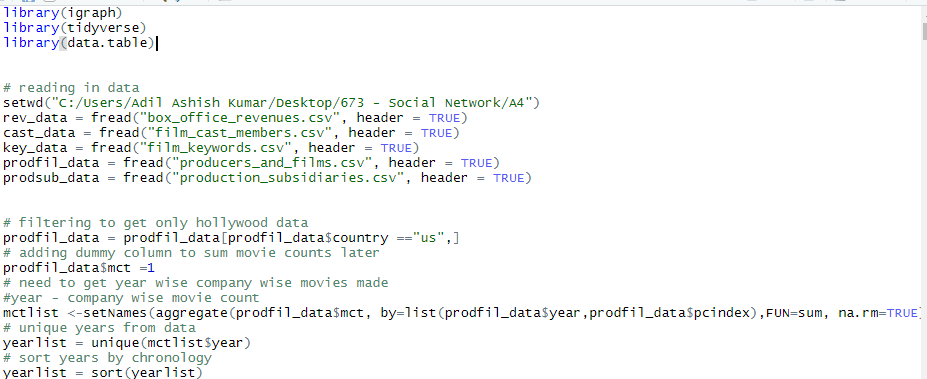
iii. Central co-productions: films made by a group of multiple generalists

iv. Peripheral co-productions: films made by a group of multiple specialists

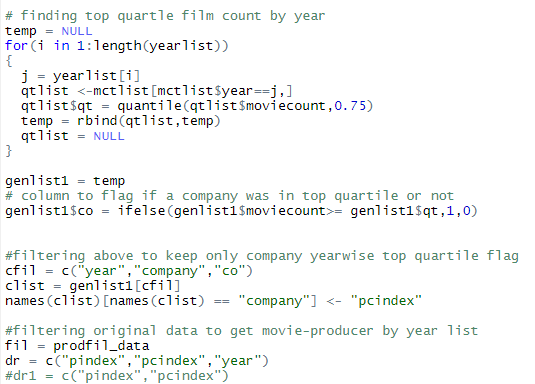
v. Hybrid co-productions: films made by a group of generalists and specialists

For each measure of generalism, a figure that illustrates the number of new keywords and new combinations of existing keywords that are introduced per type of film over the course of the data. On the *x*-axis should be years, and on the *y*-axis should be the count of new keywords or new combinations.

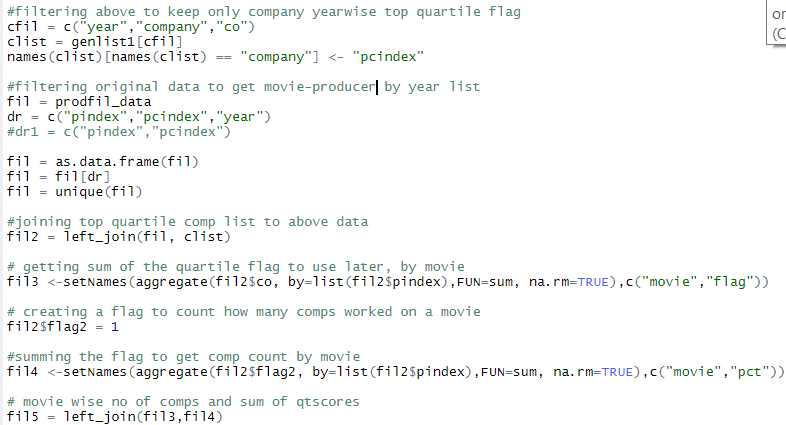
I started by loading all the data into R. I then filtered the films file to keep only Hollywood data. I added a dummy column to aid in counting no of movies by prod company for each year. I used the aggregate function to do exactly this by summing the dummy column by pcindex-year combos. I then got list of unique years in the films data and sorted and stored it in a variable



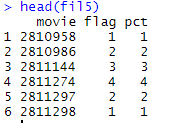
I then used a loop to loop through each year and find the top quartile of movies made in a year across all producers. I used the quantile function and specified 0.75 as an argument to give top 25 percentile of movies made. I now have year wise top quantile as a column, against which I can compare if a producer was in the top quantile or not. I used an ifelse statement to assign 1 to a column if a producer was>= the top quartile and 0 else.



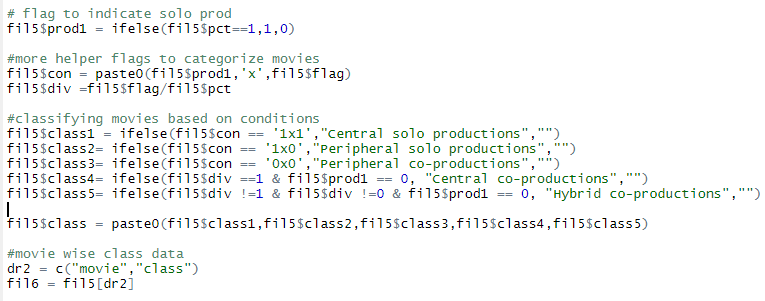
I then filtered the above dataframe (df) to get producer year wise quartile flag (if producer was in top quartile or not). I joined this to a df with producer-movie wise data to get a df with movie- producer-year wise quartile flag. I then aggregated this quartile column at a movie level to use as helper in classification. I did this using the aggregate function. I then added a dummy flag to keep track of no of producers per movie. I then aggregated this flag by movie to get movie wise count of producers, using the aggregate function.

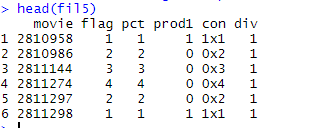


I then joined the 2 dfs to get a df with movie wise no of producers(pct) and no of generalists(flag).



Since I have a column that indicates how many producers were part of a film, I created another column based on this to indicate if a movie was a solo production or not. I then combined this column and flag( no of generalists) to create another helper column. I also divided the flag/pct, so it will tell me if there were only generalists or not per movie.





I then used all above helpers to create the requisite classes for the first measure of generalism.

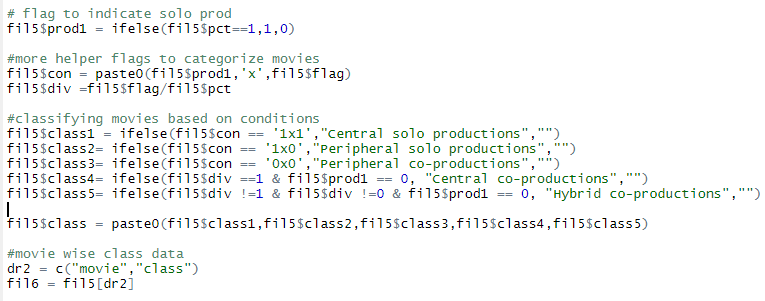
The first condition, ie class1: is picked by looking for 1x1 in the con column (1 indicating solo production , x is a separator , and 1 indicating only 1 generalist), to give central solo production.

The second condition, class2 is picked by looking for 1x0 in the con column( 1 indicating solo production , x is a separator , and 0 indicating no generalists), to give peripheral solo production.

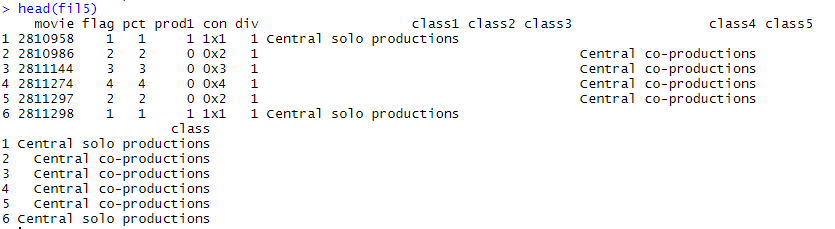
The third condition, class3 is picked by looking for 0x0 in the con column(0 indicating co production , x is a separator , and 0 indicating no generalists), to give peripheral co production.

The fourth condition, class4 is picked by looking for div=1(no of generalists = no of producers on film), prod =0 (indicating co production) to give central co production.

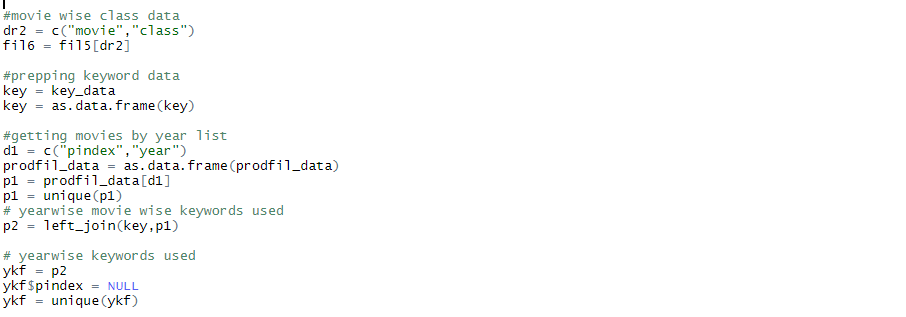
Fifth condition , class5 is picked by looking for div not equal to 1 or 0 ( no of generalists < no of producers on film), prod =0 (indicating co production) to give hybrid co production.

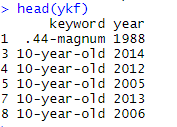


Below is resultant df with movies classified.

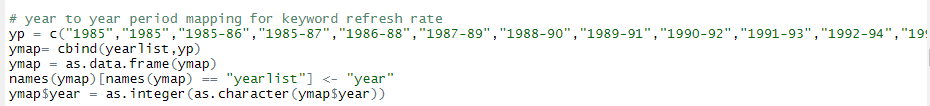


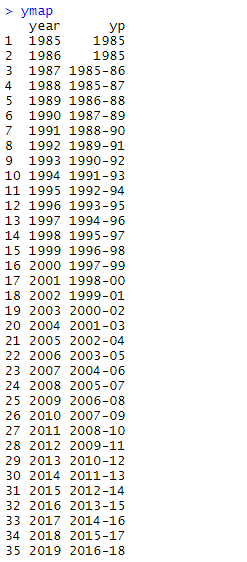
I then filtered above df to get movie wise class. I then started working on the keywords file. I filtered the films file to get year wise list of movies and joined it to the keywords file to get year wise keywords used and dropped the movie column.





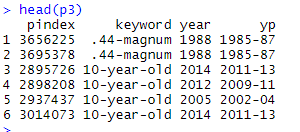
I now need to find if a keyword used was new or not. First I created 3 year groups for each year. I did this in a inefficient way( yes I know its not a very dynamic way, but it works for now, I guess), I manually created a year grouping for each year. Ex: for the year 1989, Its 3 year group would be 1986,1987, 1988. Then I can see if a keyword used in a movie in 1989 was used in last 3 years by checking all keywords in the year group 1986-88. I then joined this year grouping to a unique list of years from our films file. Now I have a df with year and its respective year grouping.



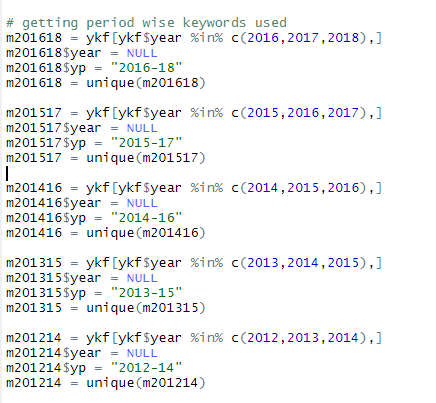


I then joined this mapping to a yearwise keyword used to get the grouping and keywords in the same df.

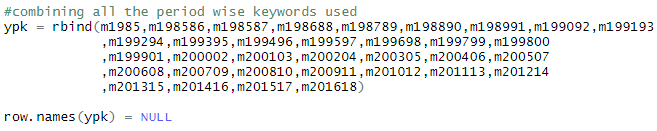




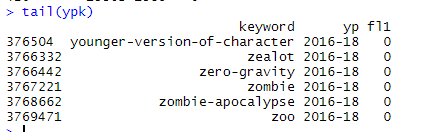
I then did another very manual/inefficient thing, by manually creating subsetting the above df for each group and then getting unique list of keywords per group used. I am aware this is not the best way to do it, but due to lack of time this is a current workaround.

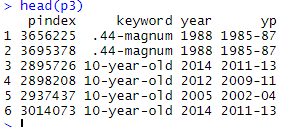


I then combined all the individual groupings into 1 df. Now I have grouping wise keywords used. I then added a flag with 0, to aid in a join.

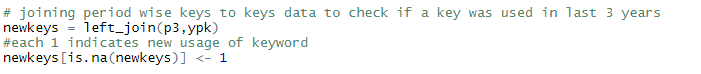






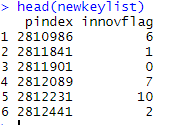


I then joined the grouping wise keywords to the df with movie- yearwise keywords used. The flag that I just defined helps to indicate if a keyword was used in its last 3 years or not. If it wasn’t used in last 3 years, then the column will have NA since it cant find it, and I replaced all NAs by 1 to indicate that the keyword is new.

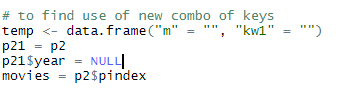


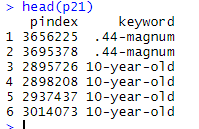
I then aggregated this flag column to get count of new keywords by movie, using the aggregate function.



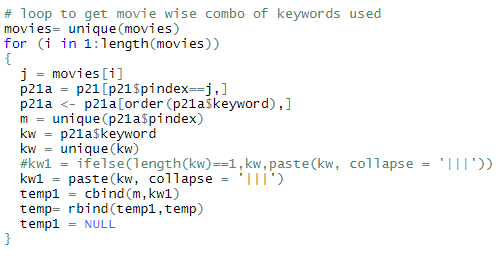


Now I need to find new combination of keywords in a movie. I take the same df I used earlier, movie wise keywords used. Then I get unique list of movies from this df.

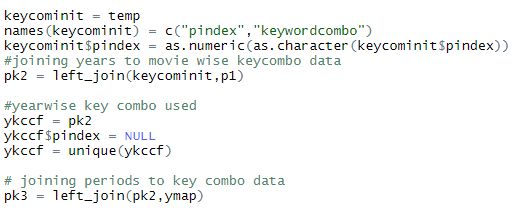


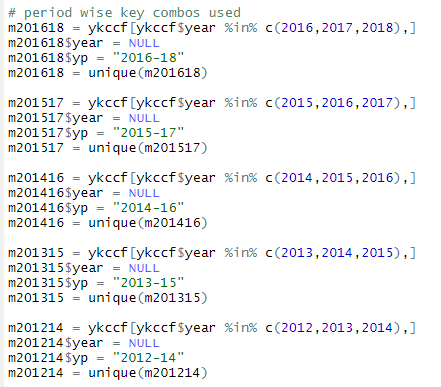


I then use a loop to loop through each movie, and calculate a column to store combination of all keywords used in that movie. I make sure to sort the keywords by alphabetical order. I use a separator “||” to separate each keyword.

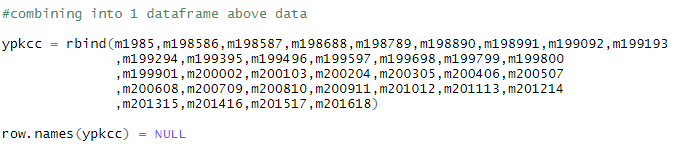


I now have a df with movie wise combo of keywords used. I then use the same logic I used to find if a keyword for a movie is new or not. I join this df to df with movie wise year release to get yearwise keyword combo used. I then use the earlier year mapping to find unique combo of keywords user per year grouping

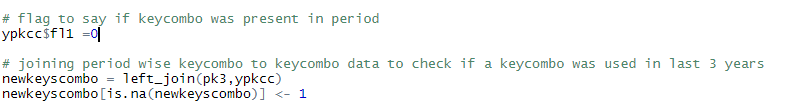




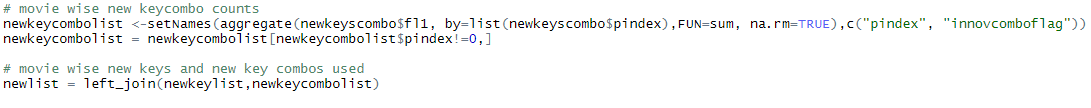
Above, I have each year grouping as a separate df, which has unique set of keyword combos used per year grouping. Ex: For 1991, I check if a keyword combo was used in last 3 years, by checking the year grouping 1988-90. I then combine all groupings into 1 df, to have grouping wise keyword combo used.

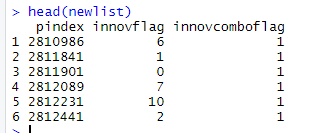


I then added a flag with 0, to aid in a join. I then join the above df to the df with yearwise key combo used. The flag that I just defined helps to indicate if a keyword combo was used in its last 3 years or not. If it wasn’t used in last 3 years, then the column will have NA since it cant find it, and I replaced all NAs by 1 to indicate that the keyword is new.

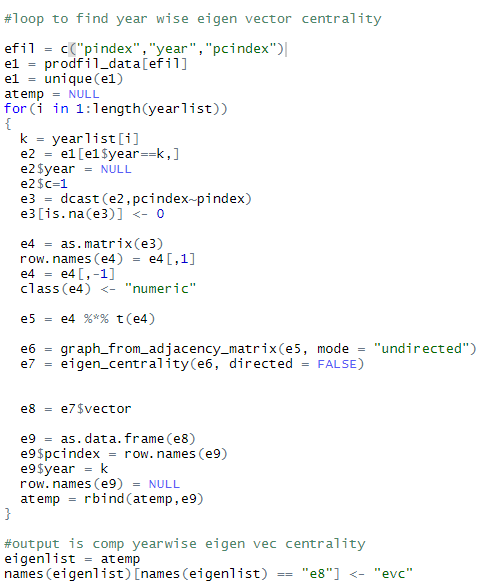


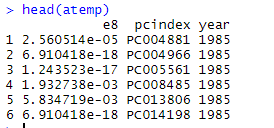
I then aggregated this flag column to get count of new keyword combos by movie, using the aggregate function. Now I have both movie wise no of new keywords and no of new keyword combos used. I join them both into 1 df.



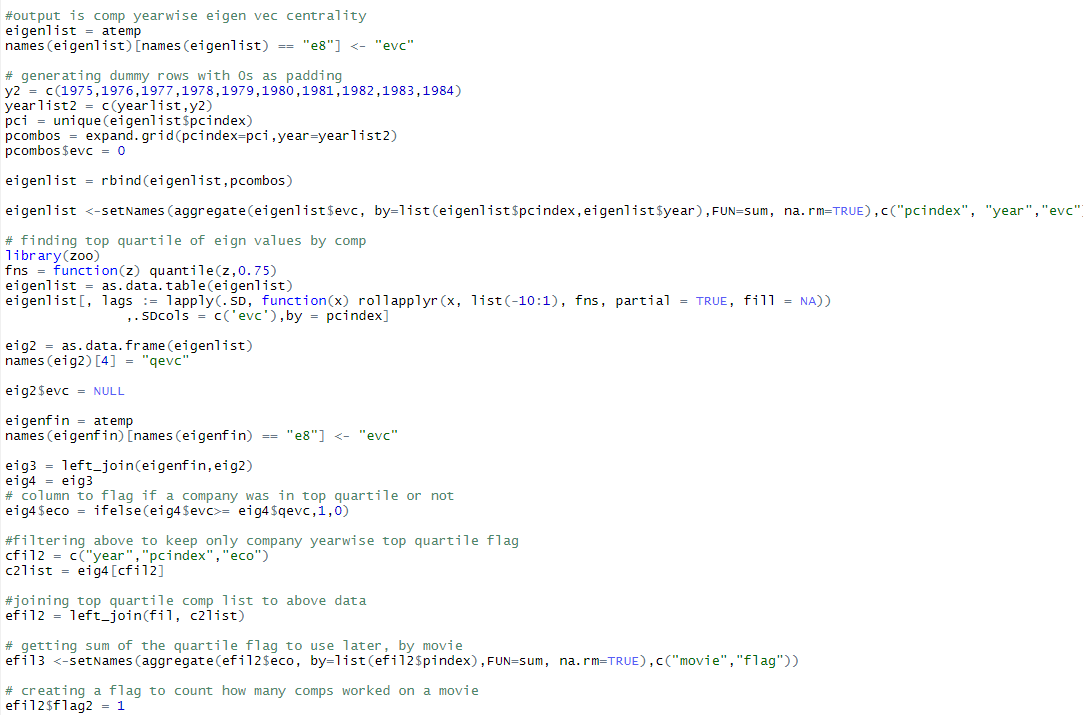


Now for the second measure of generalism, I need to calculate eigen vector centrality to decide if a producer is generalist or not. For this I started by using a df that has producer wise movie by year. Then I used a loop to loop through every year, and calculate the eigen vector centrality for each producer in the collaboration network. In the loop, I used dcast to convert a df having producer and movies into an affiliation matrix with producers as rows and films as columns. I then converted this into matrix data type and created a graph using a graph\_from adjancency\_matrix function. I used the eigen\_centrality function to to calculate the eign centrality for each producer in the network. I then converted this output into a data frame by putting the vector values and producers into a df. Now I have a df with producer wise eigen vector centralities for each year.

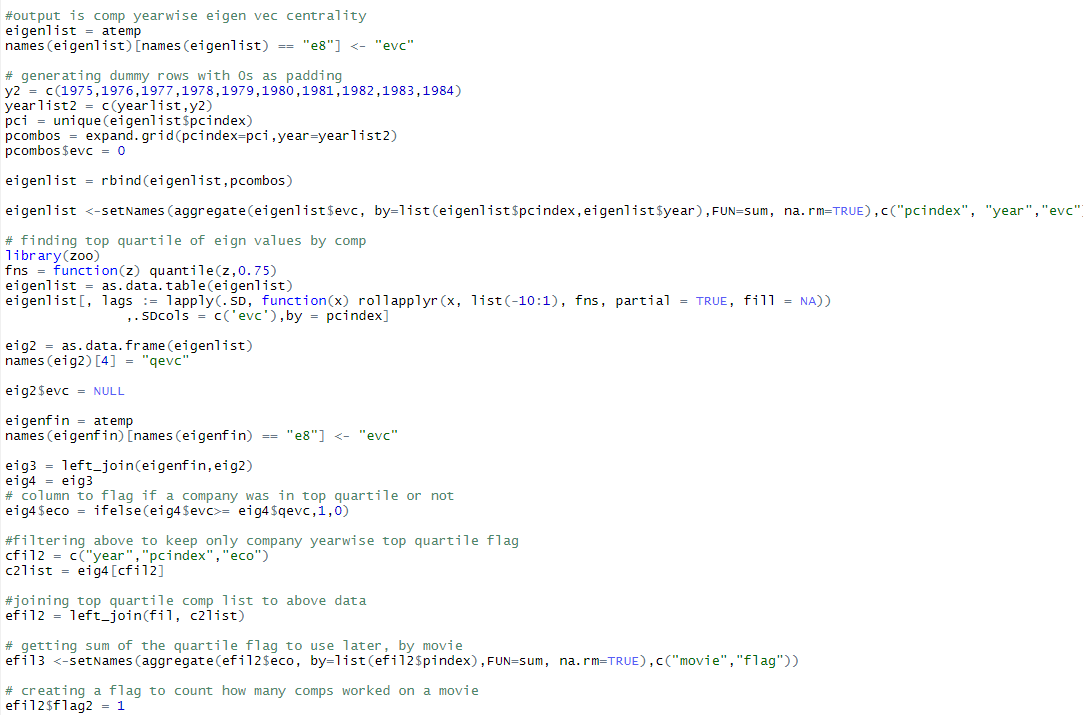


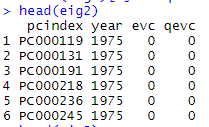


I then need to find of a producer if in top quartile of eigen centrality in its last 10 observations. First I created some dummy data with 0 values since there are cases where producers have less than 10 values. I created 2 vectors – one with unique list of years and another with unique list of producers.I used the expand.grid to create all possible combos of these 2 vectors into a df and gave eigen centrality =0 . I then used rbind to append this dummy data to original df with year-producer wise eigen vector centrality. I then aggregated this df to roll up the eigen centrality values for each company year combo, using aggregate function.

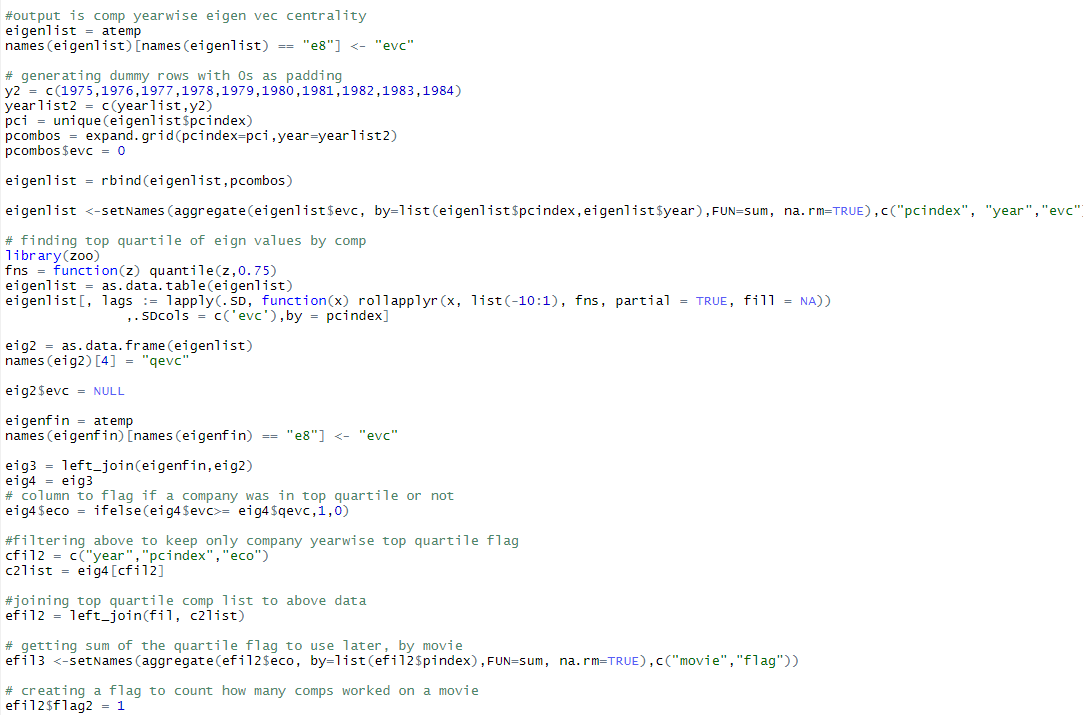


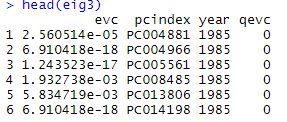
I then used rollapplyr to apply a function I created to calculate the top quartile for each year based on its last 10 observations.



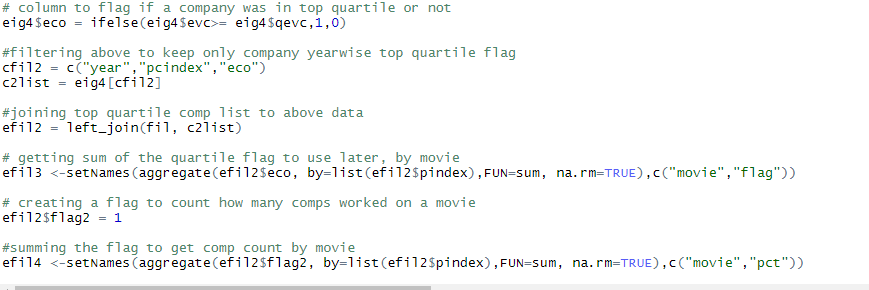


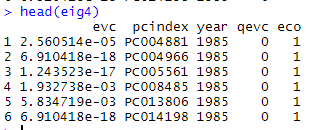
I now have top quartile eigen centrality for each year. I join this back to df with year -producer wise eigen centrality. I do this so that the dummy data is filtered out.





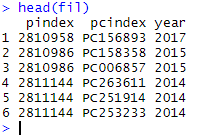
I then defined if a producer is in the top quartile or not by checking if the producers eigen vector centrality is>= top quartile centrality. I used an ifelse to do this. Now I have a column(eco) to indicate if a producer is generalist(1) or specialist(0) based on the new measure of generalism .



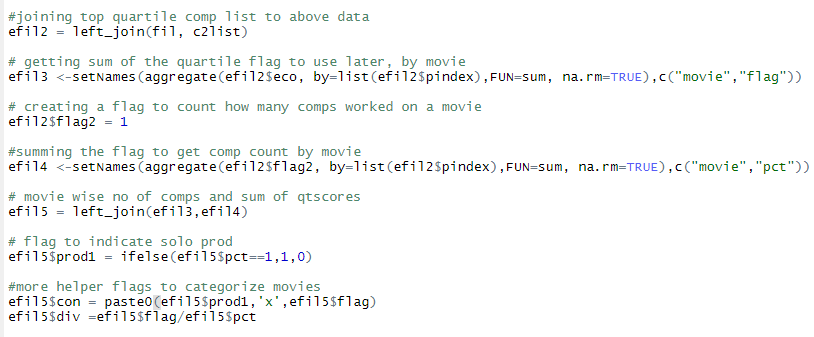


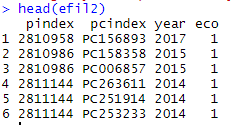
I then filter above df to get producer yearwise generalist flag.



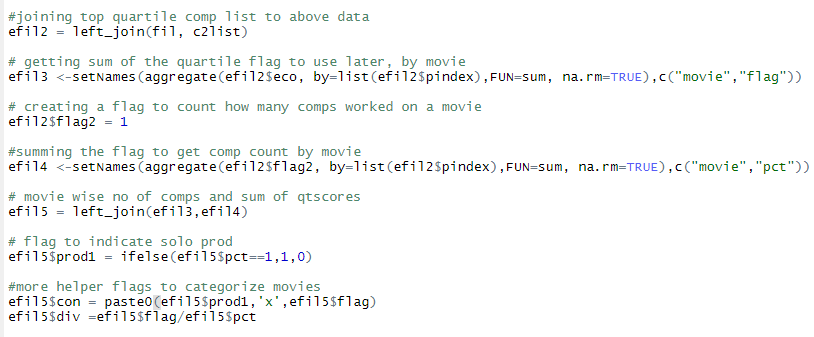


I then join the above df to producer- movie list by year. Now I have movie level generalism.

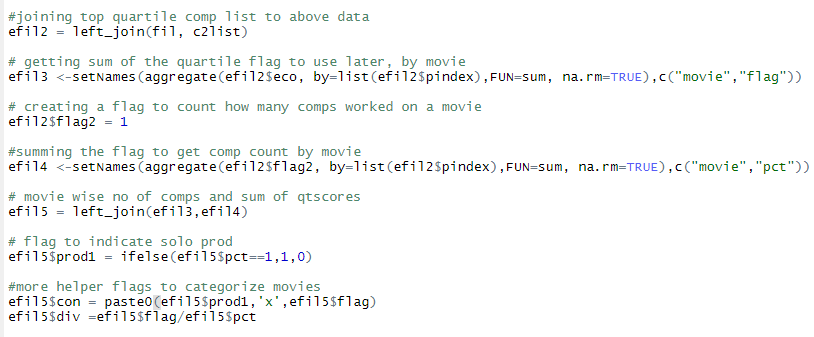




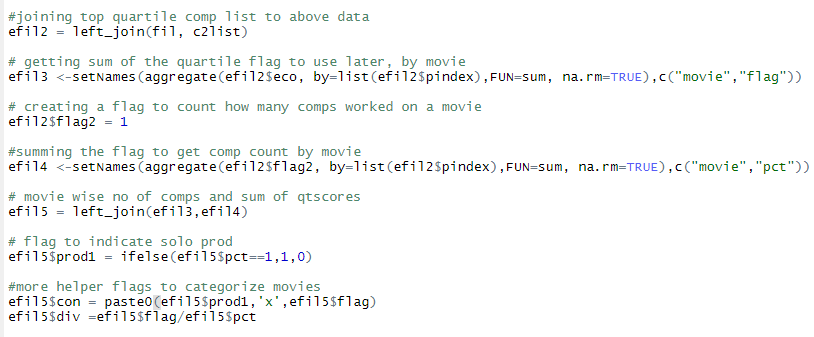
I created a flag to count producers per movie



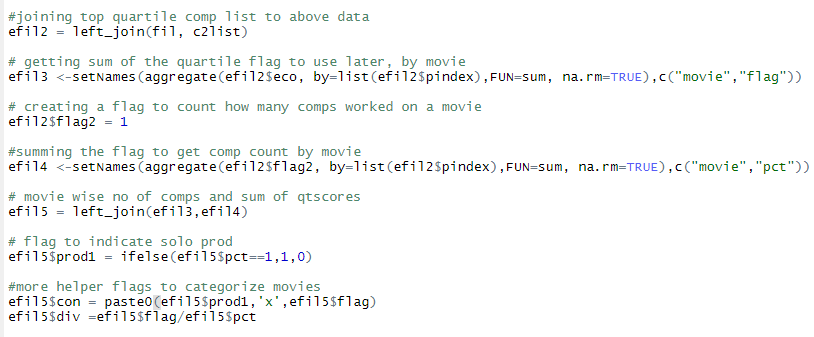
I then aggregate the generalist flag to get movie wise count of generalists. I use the aggregate function.

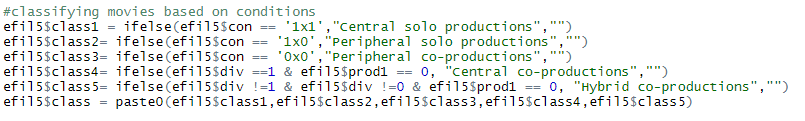


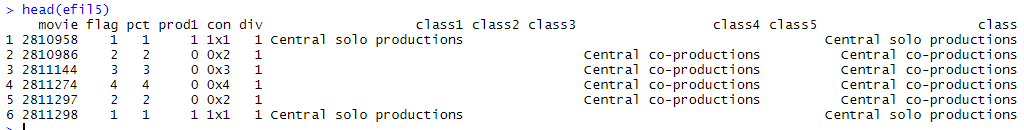
I aggregated the flag to then count no of producers per movie using the aggregate function.



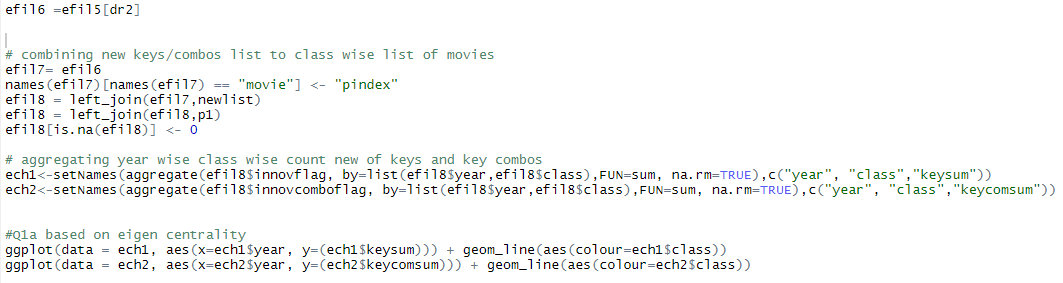
I then join the 2 dfs into 1 df to have movie wise count of generalist producers and no of producers. I added a flag to check if a movie was solo or co production. I created helper columns similar to the case of first generalism to classify a movie.

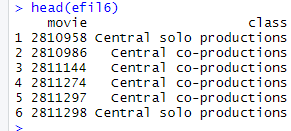




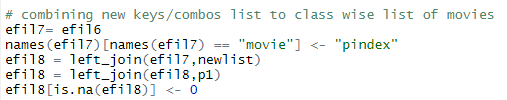


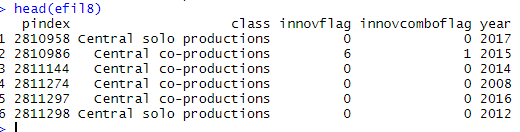
I then filtered above df to give movie wise classification.





I then joined the movie wise new key and new key combo count df to the above df. I now have movie wise class and count of new keys and new key combos.



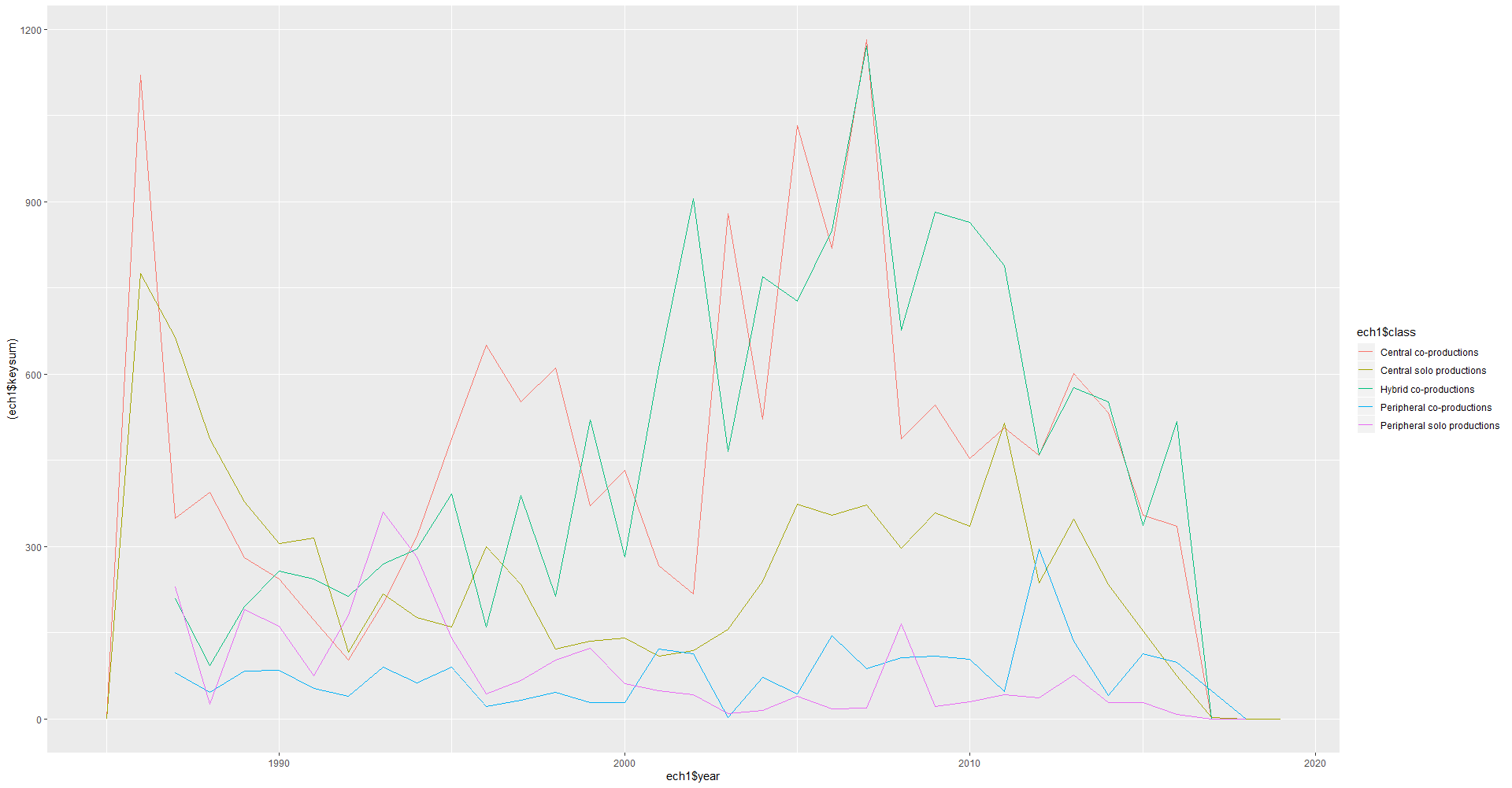


I then aggregated sum new keyword and sum of new keyword combos by year- class, using aggregate.

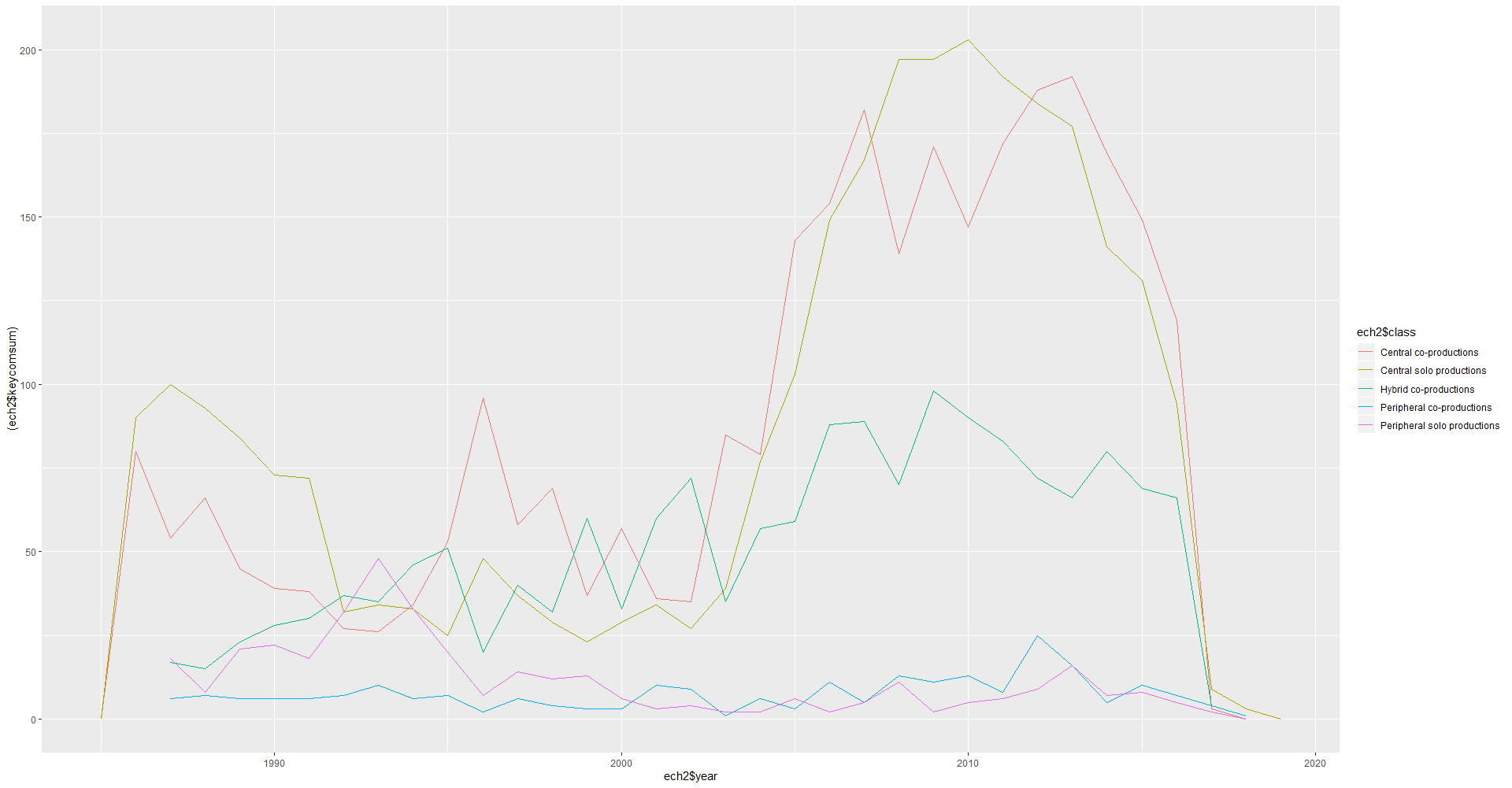


Finally, I have data ready for the plots for Q1a. This is based on eigen vector measure of generalism.

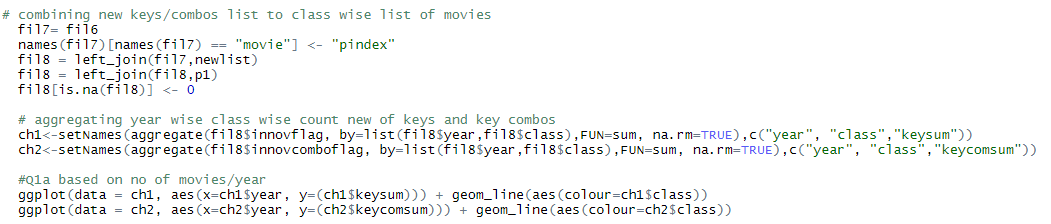
First plot shows new keywords by year

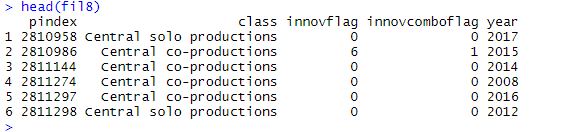


Second plot shows new keyword combos by year

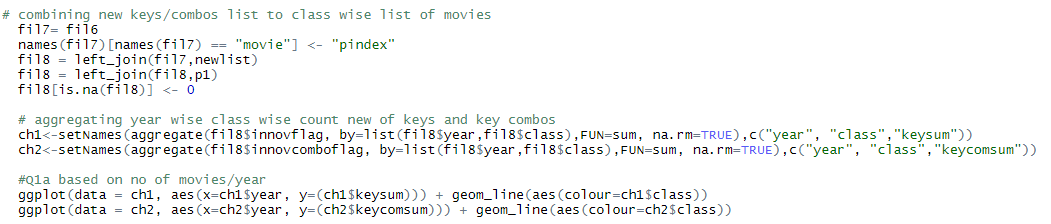


Similarly, for 1st measure of generalism (no of films per year) , I joined df containing movie wise class and keyword sums data to get movie wise class wise new keywords and new keyword combos.

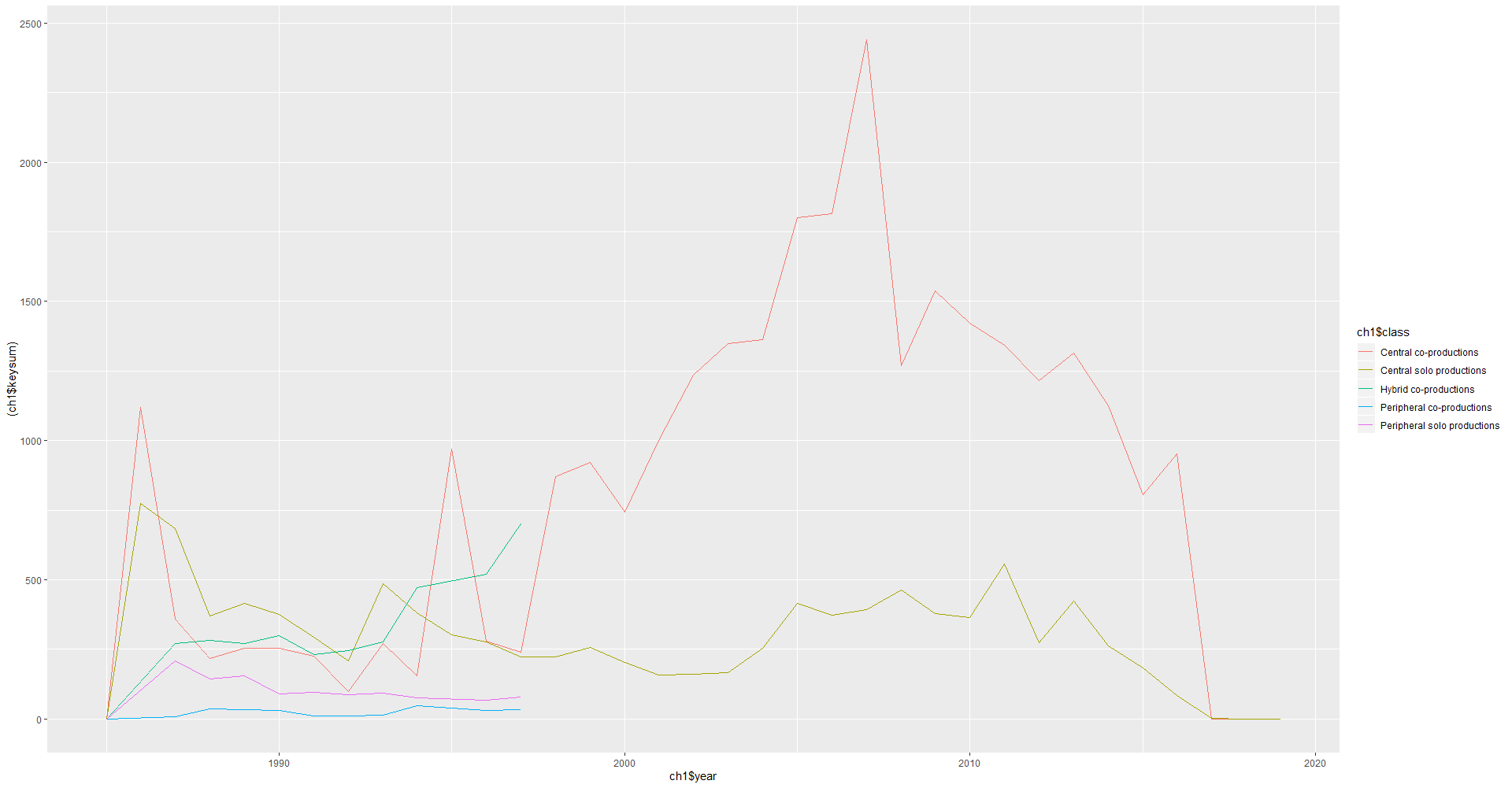




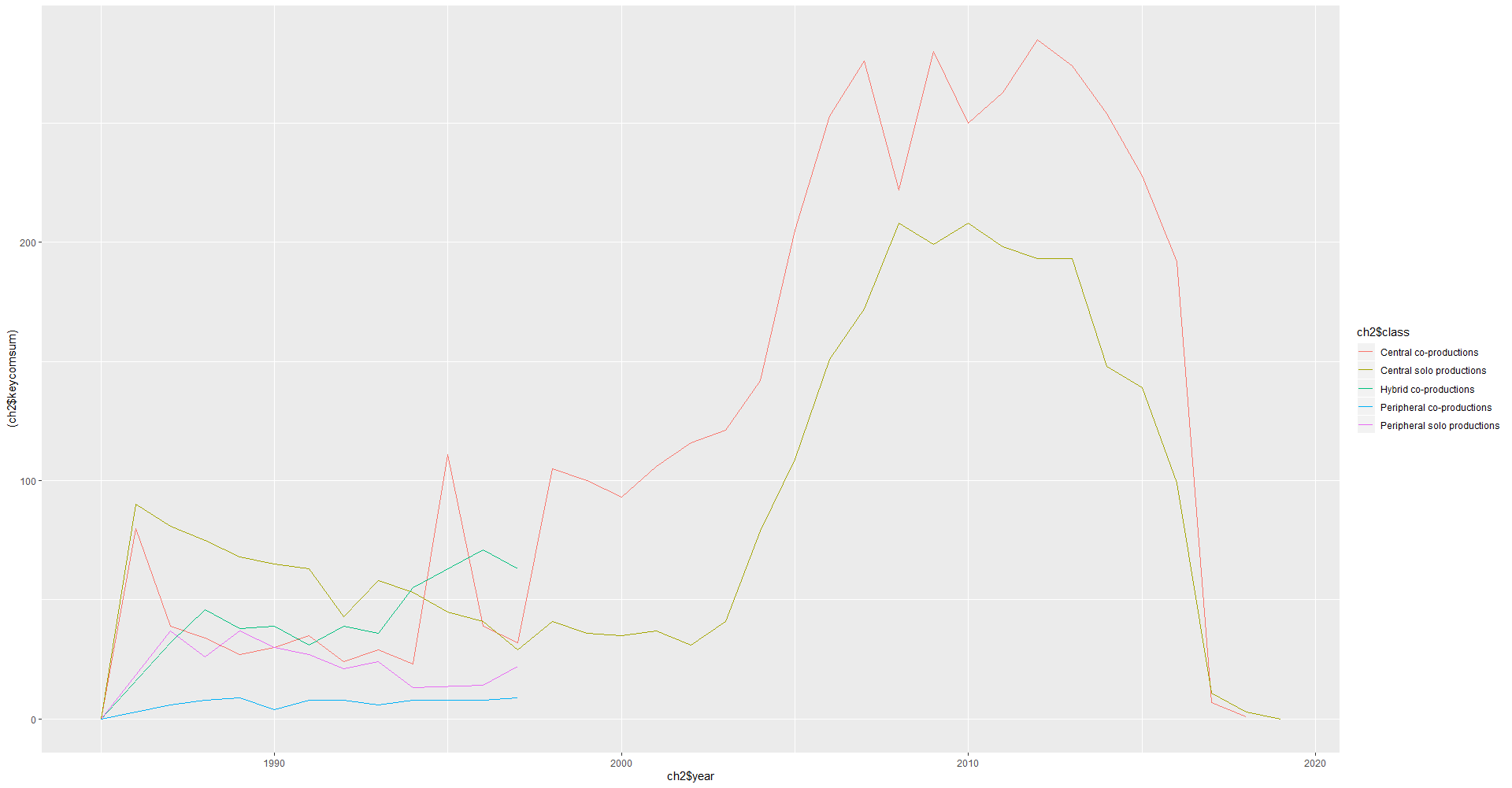
I then aggregated sum new keyword and sum of new keyword combos by year- class, using aggregate.I now have data ready for plots for 1st measure of generalism.



First plot shows new keywords by year



2nd plot shows new keyword combos by year



(B) For each measure of generalism, estimate one regression predicting the number of new keywords and another regression predicting the number of new combinations of existing keywords producers introduce in a year. Use as predictors the number of films a producer makes that year that year that fall into each of the three co-production types. So, there will be three collaboration predictors:

i. Central co-productions: number of Central co-productions a producer made that year

ii. Peripheral co-productions: number of Peripheral co-productions a producer made that year

iii. Hybrid co-productions: number of Hybrid co-productions a producer made that year

Also include control variables for a producer’s box office revenue that year, how many years the producer has been in operation, whether or not the producer is a subsidiary, and a time trend for each year.

Since it is possible for some genres of films to be more innovative than others, also control for the content of producers’ films. To do this, perform a multidimensional scaling using two dimensions that uses as the input the Jaccard distance between each producer based on the co-occurrence—the overlap—of keywords that they use in their films. To account for the natural time cycle of the production process, use as the comparison set for similarity the current year as well as the two years before the current year. You can calculate Jaccard distance using the dist() command from the proxy package, and you can perform the multidimensional scaling using the cmdscale() command from the stats package, which is automatically loaded when R starts. Use the two coordinates produced by the multidimensional scaling as controls in the regression.

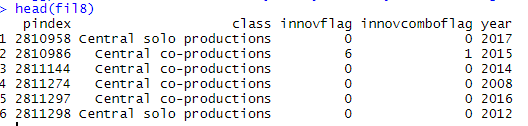
Similar to the political parties on the previous exercise, the outcome variable is a count, so we can use a regression adopted for data of this form using the MASS package, using a model specified in the form of glm.nb(new keywords variable \_ Central co-productions + Peripheral co-productions + Hybrid co-productions + Coordinate 1 + Coordinate 2 + Total box office +

of years in operation + Is subsidiary + factor(year), data, offset(total films made that year, for which there is keyword information))

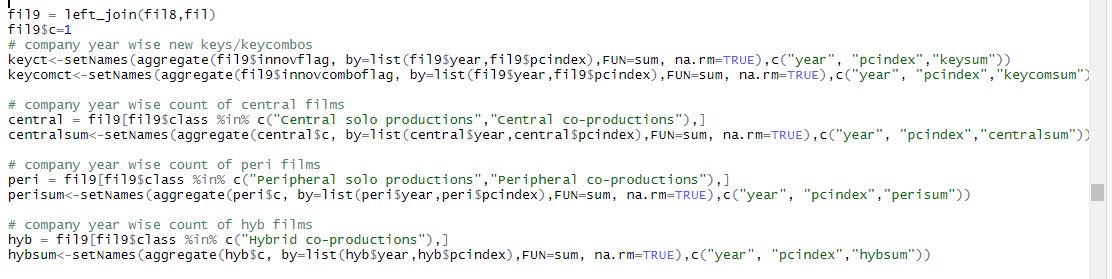
The offset accounts for the fact that making more films provides more opportunities to produce new keywords—it allows us to estimate the outcome as a per-film rate.

What kinds of collaborations seem to result in the most new keywords and new combinations of existing keywords? Comparing the two measures of generalism, are collaborations between large or small companies or core and peripheral companies more effective for creative innovation?

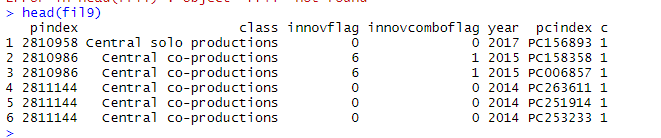
I have the below df from last question that has movie- class wise count of new keywords and new keyword combos. I join this with a df having movie wise producer list to get this data at producer level.

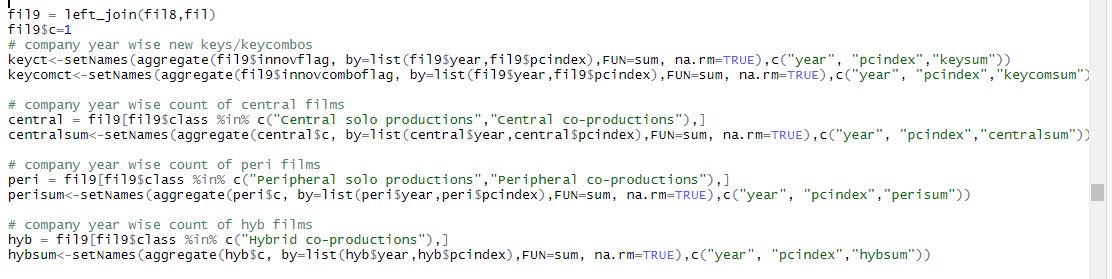


I used aggregate to get sum of new keywords and sum of new keyword combos by year-producer, using 2 separate aggregate statements.

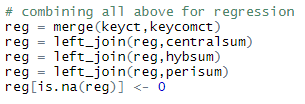


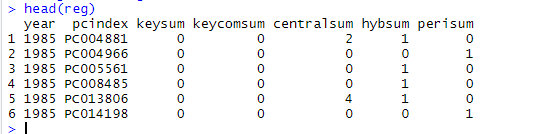
I then filtered the data to have separate dfs for central , peripheral and hybrid productions. I used aggregate for each type of production to get no of movies per type.



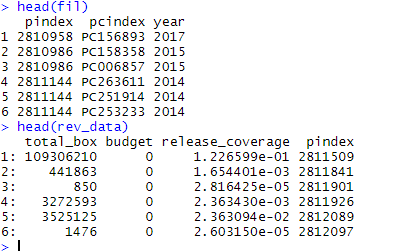


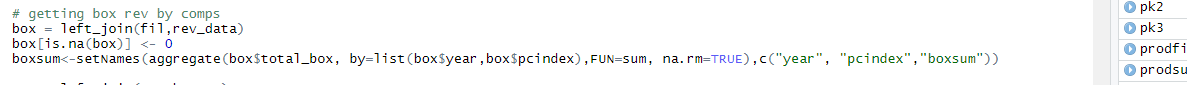
I then merged all the above dfs to get one df with all above measures.

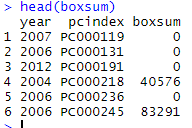




I then take the revenues data file and join it into a df having producer wise movies by year. So now I have the revenues and coverage by producer. I then aggregate box office revenue by producer-year since we need it for regression.



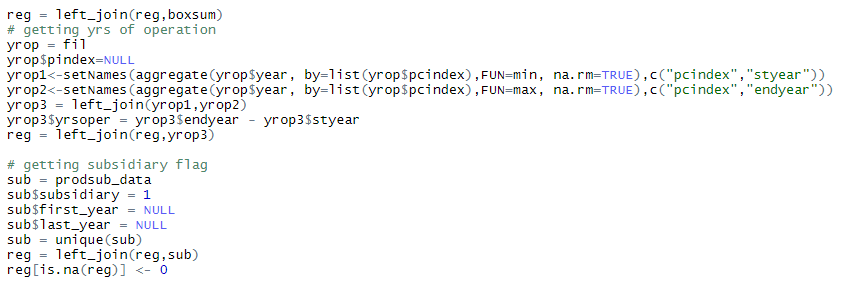




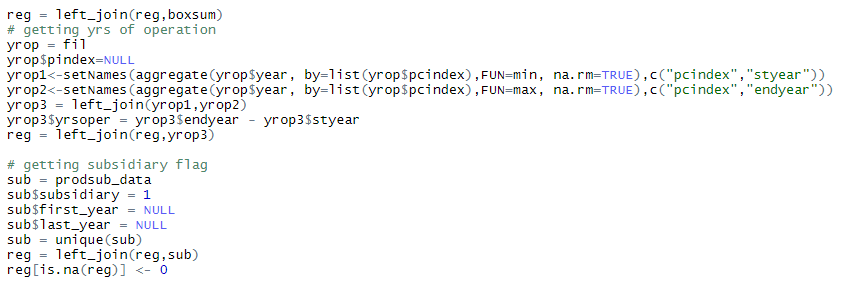
I then above df to earlier df with all variables needed for regression

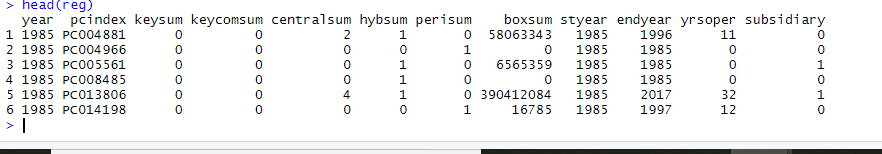


To get years of operation, I found min and max year for each producer from the films data. I did this using 2 aggregate fns, one for min and max. I then joined the 2 dfs, and subtracted min from max to get years of operation. I joined this df to the above reg df that has all variables for regression.

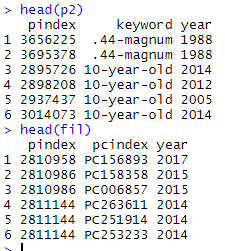


I used the subsidiary file and first added a dummy column with value of 1. I then dropped the 2 year columns, so I have producer wise subsidiary flag. I then joined this into our reg df that has all reqd variables for regression.

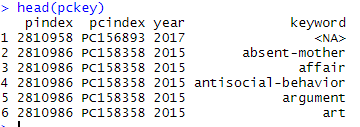




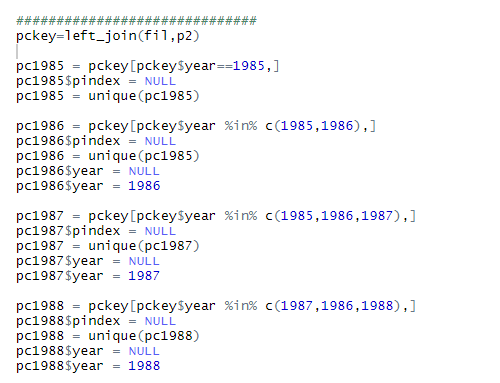
I also need to find jaccard distances and multidimensional scaling cords in the regression. I need a matrix of producers by keywords used. First I join the 2 dfs below to get producer wise keywords used.



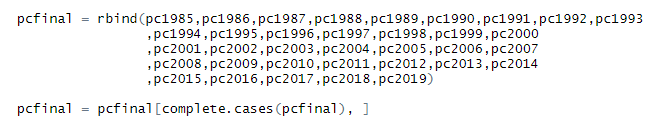




Then I manually define year groupings of 2 previous years for each year, to account for keywords used every 3 years. So I have list of keywords used by a producer for every 3 year cycle – in year 2007, I have all keywords used by keyword for 2005-07.

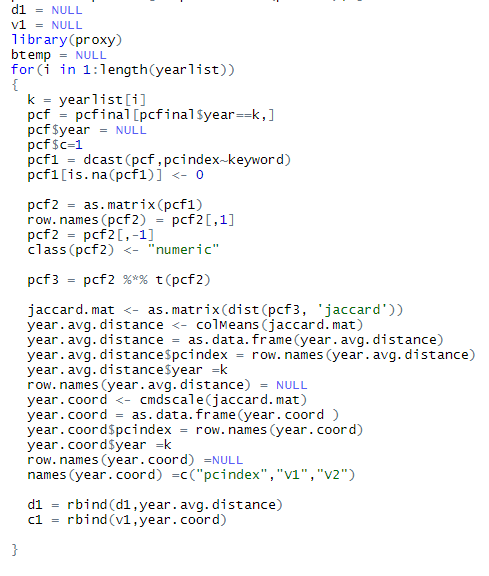


I now have keyword used by producers for every 3 year period. I now combine all these into 1 df.



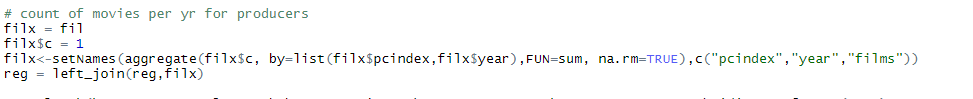
Now that I have a df with producers and keywords they used in every 3 year period, I can convert this into a matrix to find jaccard distance.

TO do this I used a loop. I looped through each year. I then used dcast to convert a df with producers and their keywords used into a matrix with producers as rows and keywords as columns. I then converted this affiliation matrix into an incidence matrix. I then used dist() from proxy package to find jaccard distance. I then took mean distance for each producer from the resulting matrix and stored it in a df. I also used this matrix of jaccard distances to calculate the multidimensional scaling and get the coordinates, using the cmdscale function.

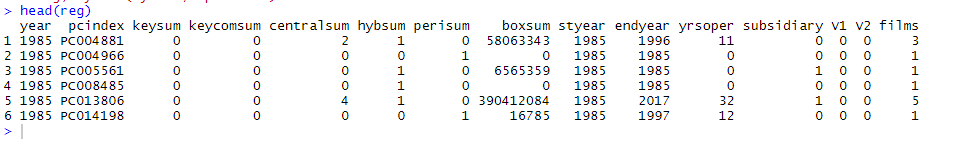


As a result of this loop , I have 2 dfs, yearwise producer wise avg distances and scaling cords.

To get count of movies per producer every year, add a dummy column to df having producer, films and year columns. I then use aggregate to count no of movies per producer every year.I join this back to reg df that has all variables for regression



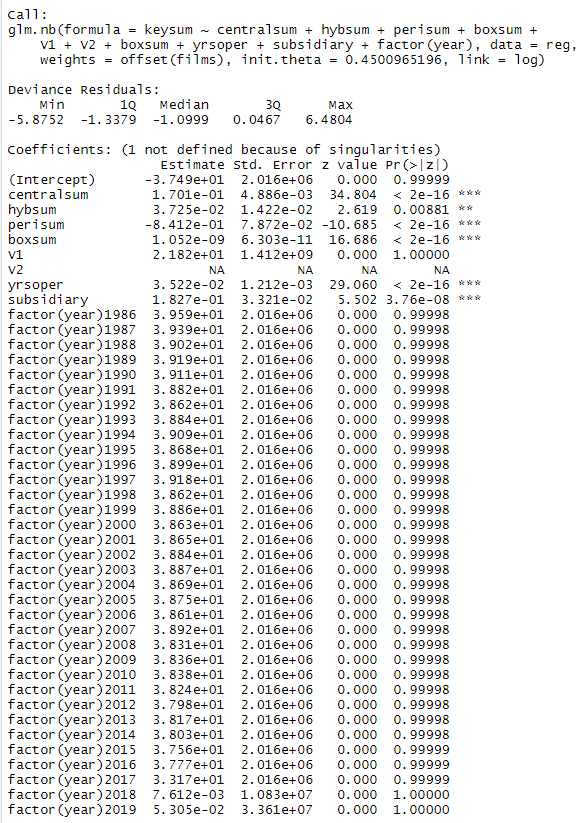
I finally have data ready for the regression.

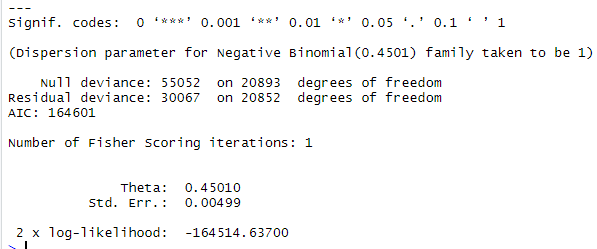




I used the glm.nb fn from MASS to do the regression to predict **new keywords** using no of central productions, no of hybrid productions, no of peripheral productions, box office revenue, years of operation, subsidiary(1 or 0) , coordinates from multidimensional scaling, years and no of films at producer level as offset. This is for the 1st measure of generalism – no of movies.

From regression output, we see that central productions and hybrid productions have the highest significance. Based on the coefficients, we can say that central productions tend to produce most new keywords, followed by peripheral and then hybrid productions. It is interesting to note that years of operation is also a very significant variable, indicating that more experienced a producer, higher new keywords are created from their movies. Being a subsidiary also is significant in the model , indicating producers who are subsidiaries produced movies with more new keywords.

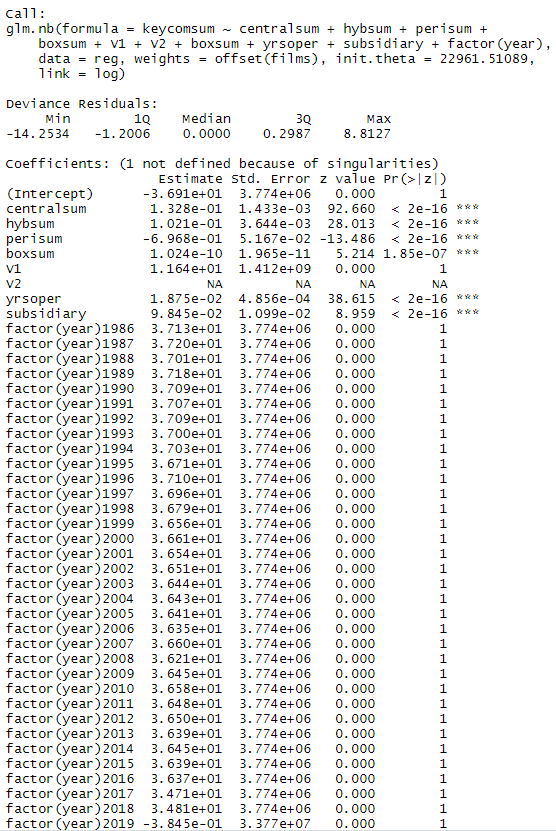


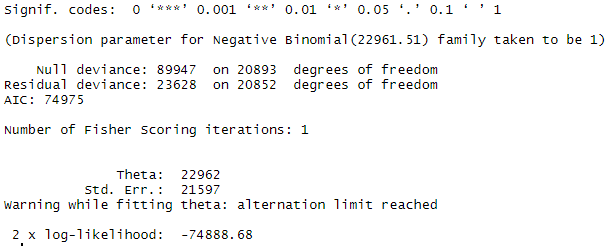


For the same measure of generalism, below is model predicting new combo of keywords



From the below regression output, we can see a similar case as for the keywords model. Central productions had a very high significance, as well as positive coefficient, indicating more central productions, more new keyword combos. The peripheral and hybrid productions also have high significance, but lower coefficient, indicating slightly lesser significance compare to central productions. Again, years of operation and subsidiary are important variables, increasing likelihood of new combos of keywords.

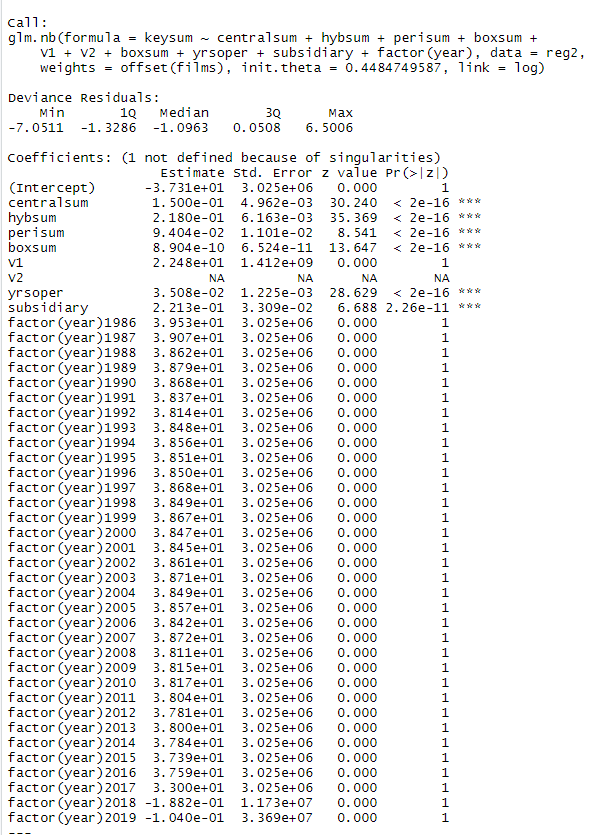


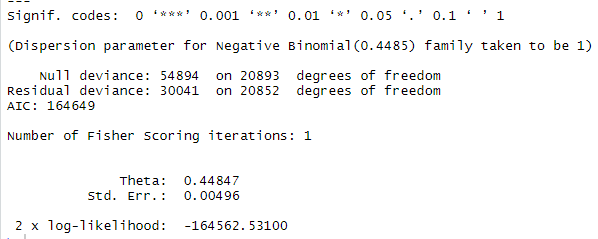


Similarly, for Eigen centrality genralism, below are the regression models

First model predict new keywords. We can see that again, all 3 types of productions are significant in model However in this case peripheral productions has the highest coefficient, indicating it is most significant in the model.Again, we can see that both years of operation and subsidiary are significant in the model

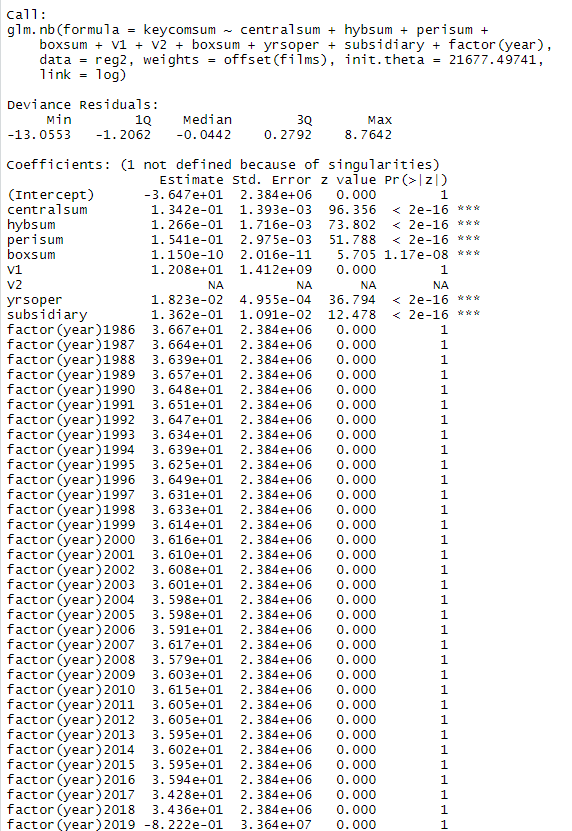


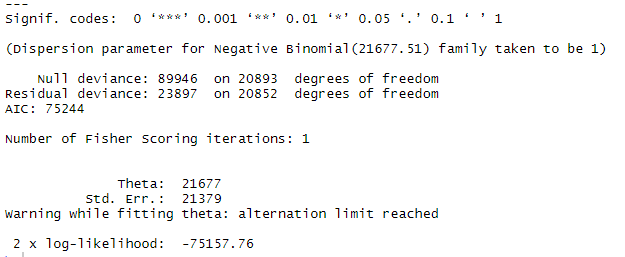




Second model predicts keyword combos. Interesting to note that all 3 production types have very similar significance in the model, with almost no difference, so all 3 seem to have equal effect on no of new keyword combos per producer. Again , years of operation and subsidiary are very significant in the model







Based on the model outputs and graphs, I would say that collaborations between firms that are core and peripheral are more effective in creative innovation. From the graph we can see more new keywords were skewed towards central firms in 1st measure of generalism, while in 2nd measure of generalism we see new keywords coming from all types of productions, indicating that collaborations are leading to creativeness.

2. What might explain why some collaborations result in more innovative films than others?

It could be that when producers collaborate with other producers that are too similar to themselves, their experience is less diverse and it is more difficult to come up with new innovations. On the other hand, when producers are too dissimilar, it can be hard to coordinate and combine very different creative ideas.

We can measure the extent to which a producer collaborates with similar producers as the average Jaccard distance between a producer and the other producers it works with based on the co-occurrence of keywords the producers use. Generate this measure yearly for each producer—again, to account for the natural time cycle of the production process, use as the comparison set for similarity the current year as well as the two years before the current year. Create a figure that illustrates how the distance between a producer and the other producers it works with relates to the number of new keywords a producer introduces each year.

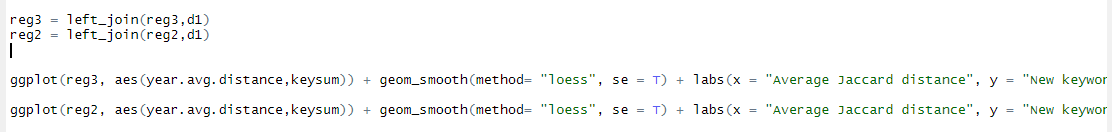
A useful way to do this is by using a “loess” smoother that plots a flexible trend line that illustrates the level of a variable on the *y*-axis at different levels of a variable on the *x*axis. Loess stands for “locally estimated scatterplot smoothing”—it fits a locally-weighted regression line over the underlying scatterplot, so it provides a tool to observe nonlinear relationships between the two variables.

You can set up a loess plot using the ggplot2 package and running a command of the form ggplot(data, aes(average Jaccard distance, new keywords)) + geom\_smooth(method = "loess", se = T) + labs(x = "Average Jaccard distance", y = "New keywords")) You can also export the plot quickly to pdf using ggplot2’s functionality ggsave("loess\_new\_keywords.pdf", width = 7, height = 7, units = "in") which lets you control the size of the pdf that is saved.

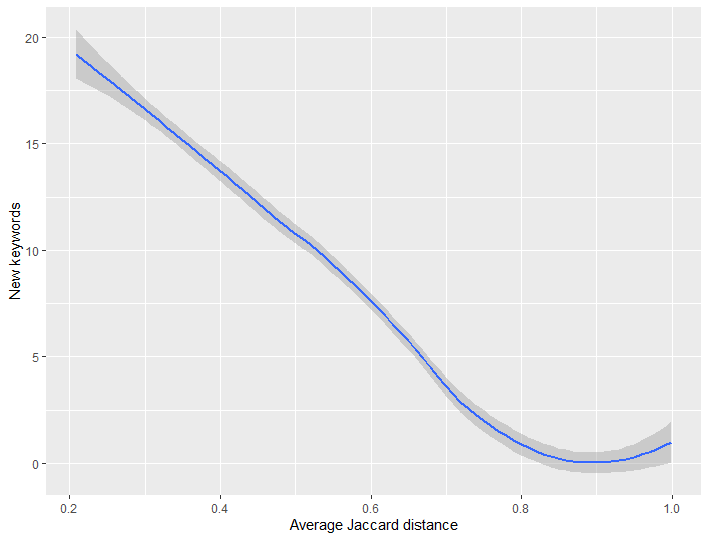
What does the pattern suggest about what kinds of collaborative partnerships might result

in more creative innovation? Does this help to explain the results from Question 1?

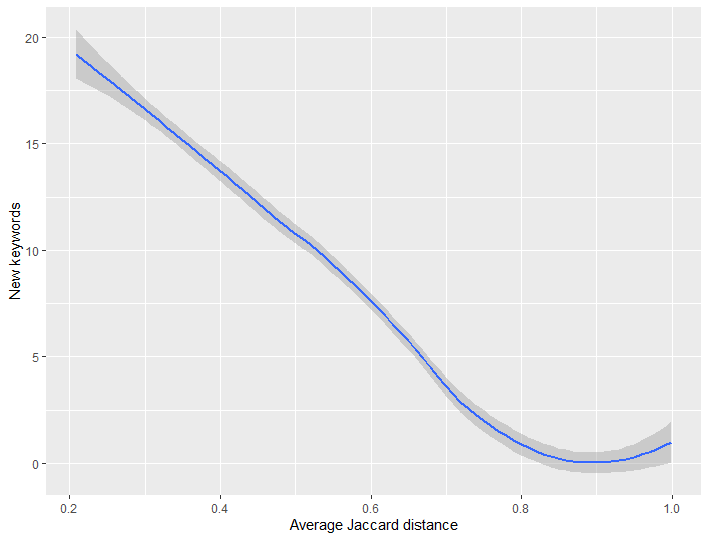
I joined the avg jaccard dist to above dfs used for regression and plotted the plots for both measures of generalism



First plot is for 1st measure of generalism



Below plot is for 2nd measure of generalism



From above plots, we see that as jaccard distance increases, no of new keywords reduce. Plots are similar for both measures of generalism. So it is not quite evident as to which measure of generalism is better for innovation and creativity.

3. Next, let’s analyze whether collaborations influence a production company’s financial returns. Since the budget information is so sparse, we will use the theater screens release coverage as a proxy for how much producers spend on each film that they make. Define each producer’s yearly return as its yearly box office revenue divided by the total release coverage it invested in for that year for its films.

To be able to make comparisons more equally across the years of the data, we’ll normalize each producer’s box office return compared to the returns that all producers earned that year. To do this, subtract the mean return of all producers for that year from a producer’s individual return and divide it by the standard deviation of the returns for all producers that year:

*standardized returnit* = *returnit* − *mean returnt*

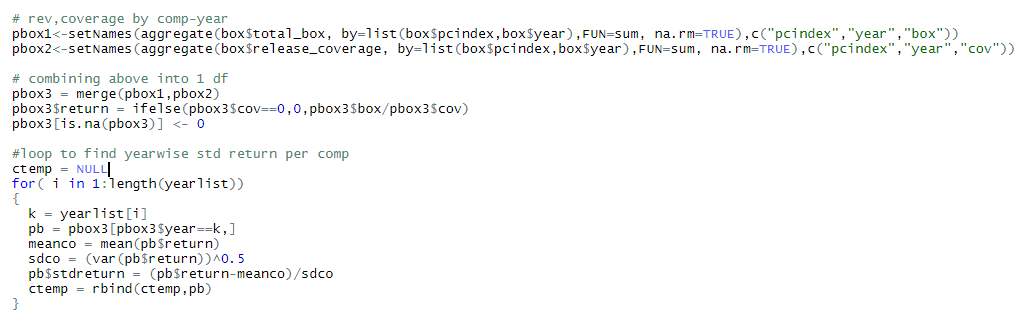
*standard deviation returnt*

for each producer *i* in each year *t*.

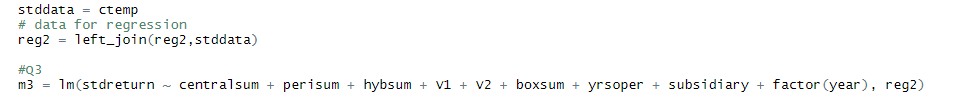
Estimate a regression predicting producers’ standardized return using the core-periphery classification to define generalists and specialists. Use the same controls as in Question 1. Since the outcome is not a count, you can estimate this model with

lm(standardized return \_ Central co-productions + Peripheral co-productions + Hybrid co-productions + Coordinate 1 + Coordinate 2 + Total box office + Number of years in operation + Is subsidiary + factor(year), data)

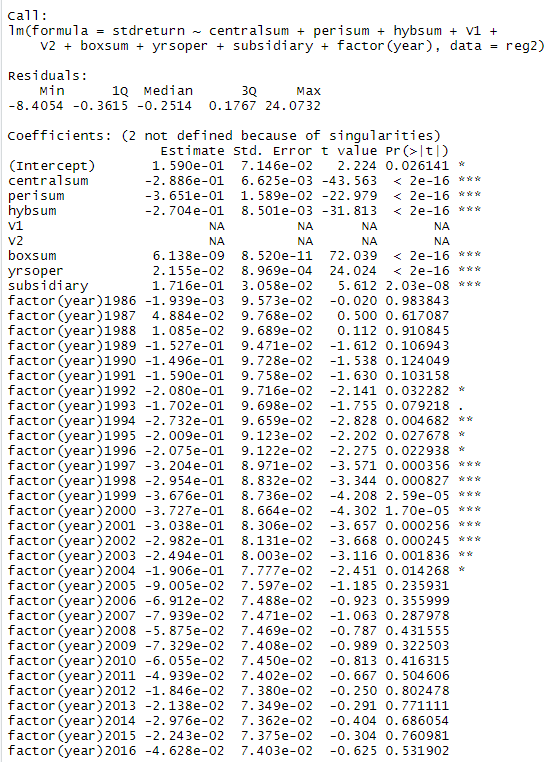
What do the results suggest about financial outcomes for collaborations?

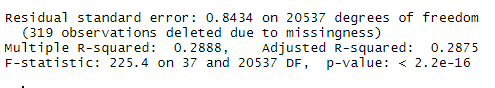


I aggregated box office revenue and coverage by producer year and merged the 2 dfs. I then divided revenue by coverage to calculate return . Used loop to calculate std return per year.



I then joined this data to data used in regression for 2nd measure of generalism. I then predicted std return based on same controls as Q1.





Model results say that all the 3 types of productions have high significance in the model – central,peripheral and hybrid. However, their coefficients are negative, indicating small decreases in return for increase in central,hybrid,peripheral productions. It also shows box office revenue is highly significant, which is understandable since return is calculated from it. Years of operation has highest coefficient among significant variables, indicating , more the years of operation of producer, the highest chances of increasing returns.

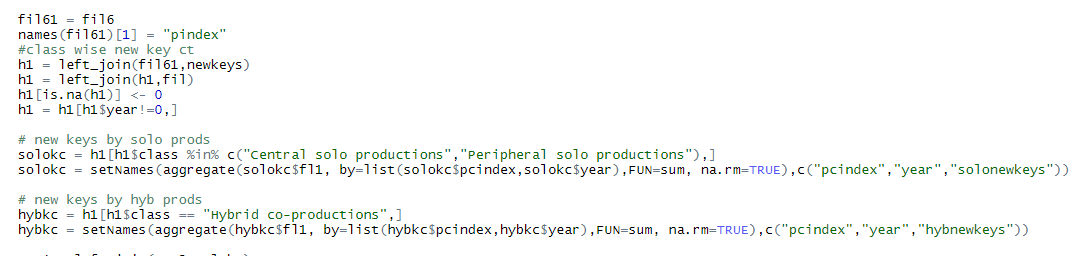
4. Collaborations can be financially risky because of the coordination required to integrate multiple producers’ experiences into a making new film. Do producers gain anything from these collaborations creatively or financially in the long term?

(A) Estimate a regression predicting the count of new keywords introduced in a producer’s *solo* produced films in a year.

Use as a predictor the cumulative number of new keywords a producer has introduced in all of its films through the current year that were made in “hybrid” collaborations. Use the same set of controls as in Questions 1 and 2. The outcome is a count, so use

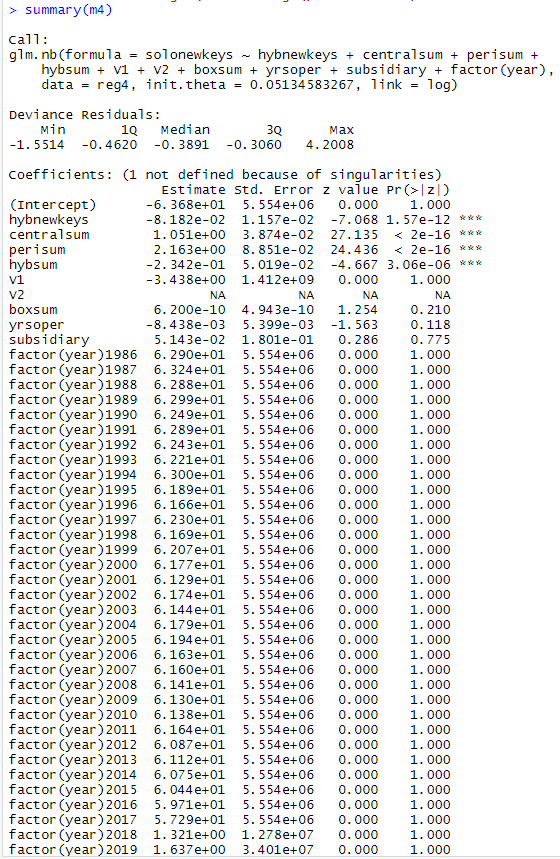
glm.nb(). Does creative innovation gained through collaborations make a producer’s solo-produced films more innovative? What does this suggest?

I used aggregate to count no of new keywords in solo and hybrid movies.



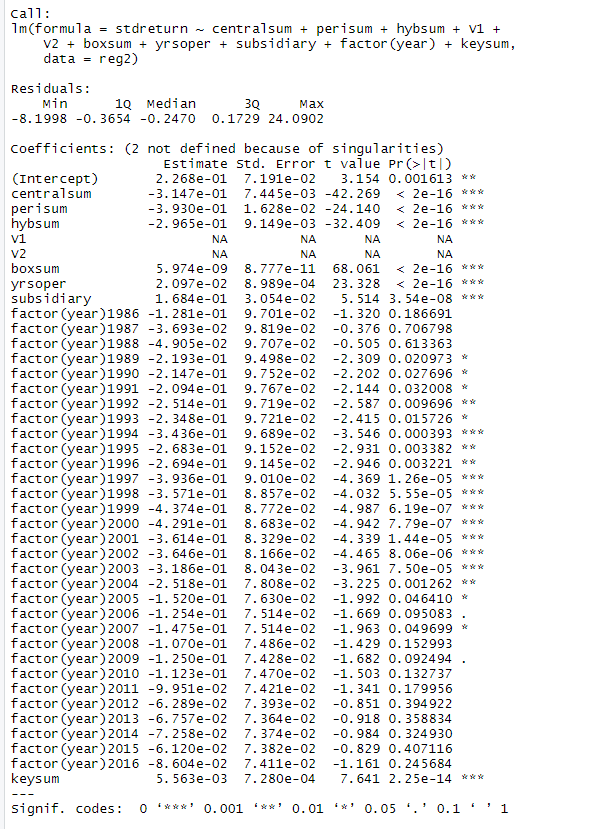
Below is regression results. We see from model that all 3 types of collaborations are highly significant in the model , indicating that creative innovation gained through collaborations does make a producer’s solo-produced

films more innovative. No other variables are significant, indicating that these collaborations are most important predictors of a producers solo movies innovativeness.



(B) Accounting for a producer’s engaging in collaborations, does introducing new keywords result in higher box office returns? To gain insight into this, estimate the same regression model from Question 2, but add in a predictor for the number of new keywords introduced.

Does this result help explain why producers might engage in collaborations, even though they can be financially risky?



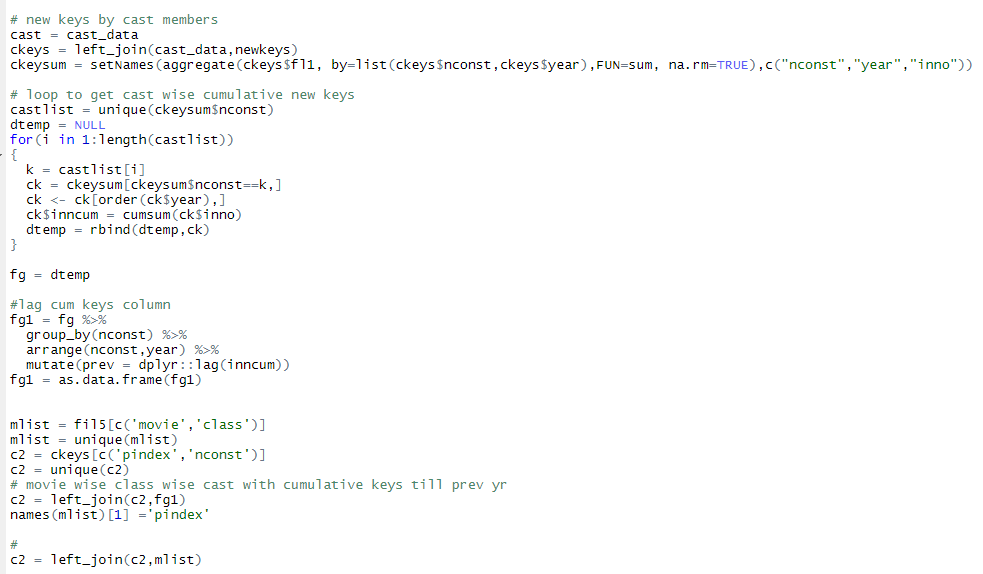
Above is the model to predict std return with added variable of new keywords. We can see that no of new keywords is very significant in the model. However the coefficient of the variable is really small, indicating small increase in returns for increase in no of new keywords.

**Extra Credit (2 points)**

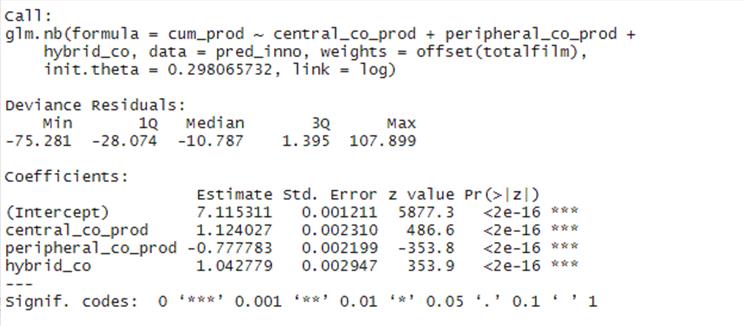
We also have information about the people that work on films as cast members. These include actors, writers, directors, and other kinds of creative talent. The file “film\_cast\_members.csv” contains a unique key identifying the creative talent and describes what role they worked on as a member of the cast of each film.

One way that production companies might benefit from collaborating with one another is that it helps them hire more innovative creative talent to work on their films.

Define a cast member’s innovativeness as the cumulative number of new keywords created in the films that they have worked on in their career up to the prior year. Using the scale-based classification to define generalists and specialists, estimate a regression predicting the innovativeness of the hired creative talent based on the types of collaborations a production company engages in. Does engaging in more hybrid collaborations seem to help with hiring more innovative creative talent?



I joined the new keywords data to the cast data file Now I have cast wise new keywords. I then used a loop to find cumulative sum of new keywords up to previous years for each cast member. I now have cast member wise cumulative count of new keywords



We see that all types of productions are significant in the model to predict return