**DAVA Project Info**

**https://piazza.com/class/irzfub0nash5s9?cid=211**

I suggest that you either finish module 3 and then start the project or start the project while going over module 3 and then go over your answers in case you want to change something. There will be no hw associated with module 3.

It is an R dataset.  Just run this command in R and it'll load into R

load('movies\_merged')

[**Jingyu Guo**](https://piazza.com/class/irzfub0nash5s9?cid=211) [18 hours ago](https://piazza.com/class/irzfub0nash5s9?cid=211)

Just want to add for the project, if used R markdown to make the pdf, one can output the R code from the Rmd using :

library(knitr)

purl("CSE6242\_Project.Rmd") # to output the Rmd to executable script

 The result is saved in the same folder with same named .R script by default -- you can indicate the output etc.

And in case if one don't want to show the R code in the pdf, add 'echo=FALSE' to the corresponding R block in R markdown.

Also there could be one problem with just running R script with source('Rscript.R") -- the plot, if not explicitly print out, won't show up:

So in Rmd, we could just ggplot(...) and the plot shows in pdf. But if the same script (say output from above method) run with R, the plot won't show up -- the only way I know to overcome this issue is to pass a handle when you create the plot and print the handle out, e.g.

figure = ggplot(...)

print(figure)

So code the plotting part this way in the Rmd as well (it will still show in the pdf).

**Q2**

there are built-ins in many of the packages for doing this kind of thing. Using the tidyr::gather() function and ggplot2's grouping abilities within the aes() function, all of the processing you need is accomplished in two lines of code.

 If you want to apply a function across each element of a vector, dataframe or list, there are convenient methods for that called apply, lapply, vapply, and similar.



[**Eric Reed**](https://piazza.com/class/irzfub0nash5s9?cid=262) [5 days ago](https://piazza.com/class/irzfub0nash5s9?cid=262)

Thanks Nick,

I am using lapply after parsing the columns.  The parsing produces lists, which seem to require a bit of handling (and seem a bit slow, but nothing serious).

**Strip off characters**

I am thinking I have to strip off some of the characters or truncate to take just the number out of the "10 mins".

The character string has both numbers and letters and only want the number for the numeric.

This seems to run fast (instantly in R Studio):

Split the Runtime column into a vector of vectors:

use space as the delimiter

each row is 1, 2, or 4 parts: ["N/A"], ["1", "min"], or ["2", "h", "30", "min"]

Apply a custom function to each runtime row:

if 'N/A', return NaN

if 4 parts, return part1 \* 60 + part3

  if 2 parts and part2 = 'h', return part1 \* 60

else return part1 #min

Update the Runtime column in the movies data frame

 strsplit() and unlist() seem to be useful ways to head down that road... perhaps after taking care of the hour/min parts.

**the instructors' answer,**

*where instructors collectively construct a single answer*

somewhere in the project I request to omit NA as little as possible. What I mean by that is if the question asks about variable x you should base it on rows where x is not missing, but avoid removing rows where other variables are missing.

**Look at box plots on a log scale**

**Q3**

no problem - TM is meant to help. If you prefer to not use the package that is fine as well.

You will need to tokenize the strings and then create a vocabulary of unique words and then for each movie find the matches in the vocabulary

**the instructors' answer,**

*where instructors collectively construct a single answer*

Let's suppose there are only three genres: action, comedy, and drama. Then you need to add to the DF three columns and for a row that is action and drama add (1,0,1) and a row that is comedy add (0,1,0). The additional columns are binary vectors or sub-vectors of the entire dataframe row.

eg

Title  Genre                          Action    Comedy    Romance

A       Romance, Comedy        0             1               1

B       Action, Comedy             1             1               0

**the instructors' answer,**

*where instructors collectively construct a single answer*

by dictionary I mean a set of all possible genres (dictionary in the NLP or text mining sense rather than hash table). In order to map genres to indices in the set of genres/dictionary you must first build a dictionary (whether you write the code to do it yourself or you use a package like tm)

[**Damien Benveniste**](https://piazza.com/class/irzfub0nash5s9?cid=328) [1 day ago](https://piazza.com/class/irzfub0nash5s9?cid=328)

It is 3 lines for me to get to the binary vectors.

I use tidyr and reshape2. I strsplit, then unnest, then dcast.

instead of converting "N/A" to NA in runtime coulmn , why not just take a subset of the data where Runtime is not equals "N/A" . like subset(data, Runtime != "N/A" ).

# P1 Q3 - Top 10

Is there a way within the ggplot function to restrict a bar chart to just the top 10 counts or do I need to create a separate date frame of those genres before plotting?

[project](https://piazza.com/class/irzfub0nash5s9?cid=349)

[**edit**](https://piazza.com/class/irzfub0nash5s9?cid=349)·[good question](https://piazza.com/class/irzfub0nash5s9?cid=349)0

Updated 1 day ago by Derek Finlinson

**the students' answer,**

*where students collectively construct a single answer*

That should do the job:

scale\_x\_discrete(limits=c("Fair","Ideal"))

<http://docs.ggplot2.org/current/scale_discrete.html>

EDIT:

actually this could do the job too:

+ xlim("Fair","Ideal", "Good")

**Prof Office Hours**

what type of visualization are you expecting?

Out of all the genres, is it 10% or 15% … you can do it with a bar graph

2nd issue: exam revenue across genres: is more tricky, you can use a boxplot to view gross revenue distribution

How does it change across genres, I would not create 1 boxplot for every genre, because you loose the ability to compare across genres

**So you need to compare multiple boxplots/histograms across the same graph**

**Arrange in 1 graph the dependency of revenue and the top 10 genres**

**Q4**

<https://piazza.com/class/irzfub0nash5s9?cid=404>

Q4: Are We CERTAIN that We Want To Compare 'Year' and Year(Released)?

**the instructors' answer,**

Sorry for the confusion - I may have already removed the mismatches when I crawled the data and joined the two sources.

If you cannot find any discrepancy between "Year" and "Release" then mention this and do not remove any rows and you will get full scores for that question.

Please ignore "Date" for now as it is indeed often missing.

x$comparedates <- (x$Year == x$Released)

Let me give you a hint.  "Year" and "Released" can be different for the same movie, and yet both can be true. Movies are released on different dates in different countries, and sometimes that could result in the same movie having release years that are different. But perhaps not too much different. That thought process led to my epiphany.  This led me from 16% (too many) to 1.6% (well less than 10%, so OK).

https://d1b10bmlvqabco.cloudfront.net/photos/i4vpibwfjnI/1423960903_35.png

[**Yeshwant Dattatreya**](https://piazza.com/class/irzfub0nash5s9?cid=330) [12 days ago](https://piazza.com/class/irzfub0nash5s9?cid=330)

Great point Rafael. Is a tolerance of 2 years good enough? I see many titles which differ by more than 3 years, but going by the names, it's unlikely that there are other movies by that very name produced in a different year.

https://d1b10bmlvqabco.cloudfront.net/photos/hq9zlda4fpiv1/1460558697_35.png

[**Rafael Espericueta**](https://piazza.com/class/irzfub0nash5s9?cid=330) [12 days ago](https://piazza.com/class/irzfub0nash5s9?cid=330)

I used a smaller tolerance. After a little web searching, it seems that the release time differences are generally something like 3 months. In fact the trend through the years has been to reduce the release lag time for movies released in different countries. So the current release time differences are measured in weeks rather than months.  But this is just one reporter's opinion!

Q5

You could melt the genres into one Genres variable and facet on that (don't forget to filter only for value == 1).

http://www.statmethods.net/management/reshape.html

[**edit**](https://piazza.com/class/irzfub0nash5s9?cid=351)·[thanks!](https://piazza.com/class/irzfub0nash5s9?cid=351)0

Updated 11 days ago by Damien Benveniste

**the instructors' answer,**

*where instructors collectively construct a single answer*

one possibility is to replicate a movie in two genres to two identical movies - one in each genre - but there are other more natural ways to do this

#1

one possible conclusion is also that there is not enough evidence so support a recommendation, or that perhaps some genres do have preferred dates while others do not

Q6

# P1 Q6: Clarification on what "these variables" refer to

In the question  
  
"Comment on similarities and differences between the user ratings of IMDb and the critics ratings of Rotten Tomatoes. Comment on the relationships between***these variables***and the gross revenue. Which of***these ratings***are the most highly correlated with gross revenue (use the R function cor and remove rows with missing values)?"

1. Does "these variables" = user ratings of IMDb and critics ratings of Rotten Tomatoes, or all of the ratings?

2. Same as above for "these ratings"

**the instructors' answer,**

*where instructors collectively construct a single answer*

1 - these variables refer to all variables related to ratings including the various IMDB and Rotten Tomatos variables.

2 - ratings are a subset of the variables in 1 that relate to rating values. For example the variable votes that count how many IMDB votes would be included in #1 but not in #2.

**the instructors' answer,**

*where instructors collectively construct a single answer*

the reading is only background - no need to summarize in the assignment. The investigation is about the relationships (the reading may be useful backgrounds as you try to interpret what you see and why it happens or why it is surprising).

2. "user ratings of IMDb and the critics ratings of Rotten Tomatoes"

Are they imdbRatings and tomatoRatings? How about others like votes, reviews, etc? Should we take them into consideration?

Yes - I encourage you to take a look at all the relevant variables including votes, etc.

[**Yeshwant Dattatreya**](https://piazza.com/class/irzfub0nash5s9?cid=391) [7 days ago](https://piazza.com/class/irzfub0nash5s9?cid=391)

So far, I can infer the following, since there is nothing anywhere on the web to indicate what all these may be. Unfortunately, it's not easy to get access to Rotten Tomatoes API either. If you agree/disagree with the below, please let me know.

tomatoMeter - on a scale of 1 to 100, a rating provided by Critics

tomatoRating - on a scale of 1 to 10, a rating provided by Critics (which begs the question, why do we have a tomatoMeter and tomatoRating?)

tomatoUserMeter - on a scale of 1 to 100, a (probably weighted average) rating provided by users

tomatoUserRating - on a scale of 1 to 5, a (probably weighted average) rating provided by users (again, not sure why there are two User Ratings)

tomatoReviews - number of critics who have provided reviews (can be taken as a proxy for how many critics voted or provided ratings)

tomatoUserReviews - same as above, but provided by users.

I hope this could help newcomers, definition of tomatoMeter:

tomatoMeter = tomatoFresh/(tomatoFresh+tomatoRotten) or tomatoFresh/tomatoReviews

still not too clear on tomatoRating

Awards and Rating

How do you think these are related? Does number of Awards influence Rating or vice versa?

I was able to come up with graphs/numbers to prove both ways. But how can we determine which one influences what? or do they feed on to each other

1. A movie that gets more awards, may lead more people to watch and higher chances of getting better rating. I am more inclined towards this relationship than the next one.
2. A movie that already has high rating, gets more awards ( may be it created so much buzz that it influenced award committee )

How do we go about determining dependency between such variables which may influence/feed on each other

I hope this is interesting enough to have a good discussion.

**the students' answer,**

*where students collectively construct a single answer*

[Actions](https://piazza.com/class/irzfub0nash5s9?cid=449)

Remember - Question 6 asks about correlation, not causation.

Your questions are very interesting, but I think they're outside of the scope of Project 1.

we are allowed to use other packages like PerformanceAnalytics to show the graph of correlation

**I use (GGalley -> ggpair) I could specify for each pair, use**

**cor(..., use="na.or.complete")**

**Or, in another implementation, what i did is I melted everything, so it would be something like**

**Gross, variable, value**

**1, tomatoRating, 1**

**2, tomatoRating, NA**

**And then remove the NA in value and do ddply on variable and cor on Gross and value. This way, you also keep whatever you could**.

 I use cor(use="na.or.complete") with ggpairs

 Running ggpairs *is* computationally intensive, but here are some tips to make it work if you want:

* Pay attention to NA and N/A.
* Pay attention to column types.
* Reduce the number of columns.
* Remove all local R variables one way or another (i.e. it might not be enough to just run

rm(list=ls(all=TRUE))

,you might need to reboot!

[**Brent Wagenseller**](https://piazza.com/class/irzfub0nash5s9?cid=462) [2 days ago](https://piazza.com/class/irzfub0nash5s9?cid=462)

The way I did this was use ggpairs and simply get the correlation for each pair from the graph.  
  
It took about 2ish minutes to run, and I could absolutely not get the PDF to open, so I used PNG instead and that ran in a reasonable time and opened instantly.

[**Brent Wagenseller**](https://piazza.com/class/irzfub0nash5s9?cid=462) [2 days ago](https://piazza.com/class/irzfub0nash5s9?cid=462)

Just a note that the categories will be mushed into each other using ggpairs and I am not sure there is a way to mess with the axis (like there usually is).  I just ended up assigning a letter to each item and doing the lookup manually when I looked at the graphs.

[**Damien Benveniste**](https://piazza.com/class/irzfub0nash5s9?cid=462) [1 day ago](https://piazza.com/class/irzfub0nash5s9?cid=462)

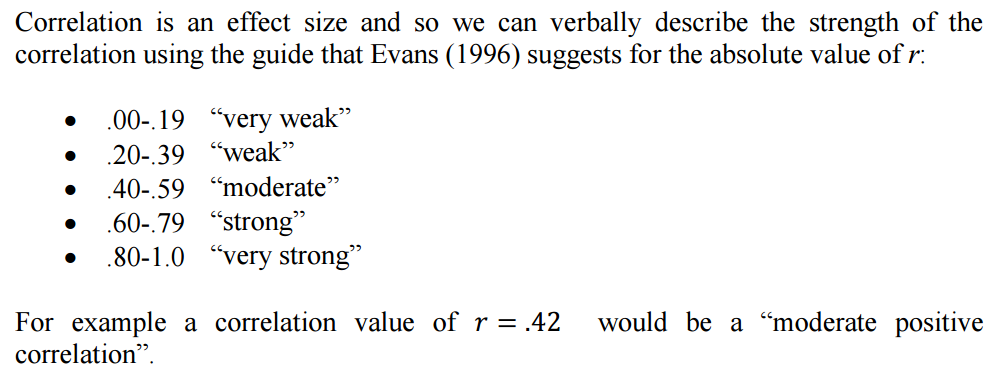
The question is asking about the ratings from imdb and rotten tomatoes. For me the way I eliminated some:

Metascore: not imdb nor tomato -> not using it  
imdbRating:**using it**  
imdbVotes: not a rating -> not using it  
tomatoMeter: **using it**  
tomatoImage: not a rating -> not using it  
tomatoRating: **using it**  
tomatoReviews: number of votes, not a rating -> not using it  
tomatoFresh: not a rating -> not using it  
tomatoRotten: not a rating -> not using it  
tomatoConsensus: not a rating -> not using it  
tomatoUserMeter: **using it**  
tomatoUserRating: **using it**  
tomatoUserReviews: number of votes, not a rating -> not using it

So I am just down to 5 of those variables.

# Q6: What is the Threshold for a Statistically Significant Correlation?

Just in case anyone else was wondering the same, I came across [this resource](http://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf) after Googling:



 Also I remind everyone to look at the low correlation too (0-.19).  You may be surprised at what is not correlated that you would think is.

[**Brent Wagenseller**](https://piazza.com/class/irzfub0nash5s9?cid=469) [1 day ago](https://piazza.com/class/irzfub0nash5s9?cid=469)

(pssst, try all of those imdb variables)

**Q7**

**the instructors' answer,**

*where instructors collectively construct a single answer*

Each row should correspond to three dimensional binary vector and yes the entire collection or rows would correspond to a n x 3 structure that can be added to the original DF

**the instructors' answer,**

*where instructors collectively construct a single answer*

looking at the data, it looks like "N/A" is either 0 nominations or not available. For the purposes of this exercise you can interpret it as 0 nomination

[**Madhav Srinath**](https://piazza.com/class/irzfub0nash5s9?cid=358) [5 days ago](https://piazza.com/class/irzfub0nash5s9?cid=358)

The point of the instructor telling us that is to give us a good way to come up with a "many" threshold ourselves while keeping that 5:1 ratio between many and some. Yes this will require some massaging of the thresholds but that's more realistic anyway.

# P1 Q7: What is being asked?

If I understood this correctly, we are asked to

1. Treat all N/As as 0 nominations + awards ([@358](https://piazza.com/class/irzfub0nash5s9?cid=358))

2. Simply sum the nominations and awards (treat them as equal)

3. Create 3 columns, "Zero", "Some", "Many"

4. Based on the number of awards and nominations, if they are not 0, split the Titles between (or assign the Title to) "Some" and "Many", so that "Some" has approximately 5 times as many rows (or 1s under it) as "Many".

[**Romeo Cabrera**](https://piazza.com/class/irzfub0nash5s9?cid=421) [5 days ago](https://piazza.com/class/irzfub0nash5s9?cid=421)

I understood that as: Let's say you mark rows with awards between 1 and 5 and "some", and rows with awards > 5 as many. Then, the ratio between the number of "some awards" rows, and "many awards" rows should be 5:1

Pd: you could use the quantile() function to distribute the date.

**the students' answer,**

*where students collectively construct a single answer*

[Actions](https://piazza.com/class/irzfub0nash5s9?cid=453)

I would make the binary vector (as this is what is requested) and then tack on a factor in addition, if it's useful to your analysis. I suspect that the binary vectors are going to be used in part 2 with some sort of regression model.

summing the wins and nominations

What I did was use str\_extract\_all to pull all the numbers out into a list, then I just summed the list.

[**Miguel Morales**](https://piazza.com/class/irzfub0nash5s9?cid=471) [1 day ago](https://piazza.com/class/irzfub0nash5s9?cid=471)

Careful, though, you might have a list of list of numbers... So summing up has to be done across the inner lists.

[**Ravindra Kumar Yadav**](https://piazza.com/class/irzfub0nash5s9?cid=471) [1 day ago](https://piazza.com/class/irzfub0nash5s9?cid=471)

Right, I handled that with unlist

[**Ravindra Kumar Yadav**](https://piazza.com/class/irzfub0nash5s9?cid=471) [1 day ago](https://piazza.com/class/irzfub0nash5s9?cid=471)

I am able to extract to list but stuck with apply/sapply for sum. Probably missing something trivial while using the construct.

I keep getting :

dim(X) must have a positive length

https://d1b10bmlvqabco.cloudfront.net/photos/hyx9thiqa6j4nn/1408295901_35.png

[**Miguel Morales**](https://piazza.com/class/irzfub0nash5s9?cid=471) [1 day ago](https://piazza.com/class/irzfub0nash5s9?cid=471)

Watch also the resulting lists contain strings not integers...



[**Ravindra Kumar Yadav**](https://piazza.com/class/irzfub0nash5s9?cid=471) [1 day ago](https://piazza.com/class/irzfub0nash5s9?cid=471)

Edit : Nevermind, I was applying intermediate functions inside apply , before passing it to sum.

I moved all processing to a custom function and called it from apply.

[**Ravindra Kumar Yadav**](https://piazza.com/class/irzfub0nash5s9?cid=471) [1 day ago](https://piazza.com/class/irzfub0nash5s9?cid=471)

I preferred strsplit(x, <regex>), where regex is anything which is not a number. This gives you list of numbers in a string.

[**Brent Wagenseller**](https://piazza.com/class/irzfub0nash5s9?cid=471) [20 hours ago](https://piazza.com/class/irzfub0nash5s9?cid=471)

Since you have the binary vectors, have you considered simply using a ddply()?



[**Laurie Ohan**](https://piazza.com/class/irzfub0nash5s9?cid=471) [20 hours ago](https://piazza.com/class/irzfub0nash5s9?cid=471)

I have:

https://d1b10bmlvqabco.cloudfront.net/attach/irzfub0nash5s9/is3mq5xtj74506/iu383n5pobqb/Capture_awards.PNG

I haven't been able to convert them to binary vectors as of yet, I had been trying with the if elses with no success

https://d1b10bmlvqabco.cloudfront.net/photos/hvil6y6khgk2ri/1408396390_35.png

[**Brent Wagenseller**](https://piazza.com/class/irzfub0nash5s9?cid=471) [20 hours ago](https://piazza.com/class/irzfub0nash5s9?cid=471)

Is that for one line? try an unlist() on it, which will force it into a single list.  from there you can simply put that in a sum()

https://d1b10bmlvqabco.cloudfront.net/photos/hvil6y6khgk2ri/1408396390_35.png

[**Brent Wagenseller**](https://piazza.com/class/irzfub0nash5s9?cid=471) [19 hours ago](https://piazza.com/class/irzfub0nash5s9?cid=471)

One other caveat would be to make sure whatever is returned in the unlist() is a number; you may have to wrap that in an as.numeric(as.character(WhateverUnlistReturns))

**Q8,**

**the instructors' answer,**

*where instructors collectively construct a single answer*

please come up with the insights that are not part of the answers to the previous questions. Given 1 or 2 or more variables you may have a preconception about what to expect. If your data supports the preconception this is an expected insight. If the data contradicts your previous conception it is unexpected. Try to come up with interesting insights. You may have to play with the data somewhat to find good ones.

[**Eric Reed**](https://piazza.com/class/irzfub0nash5s9?cid=462) [1 day ago](https://piazza.com/class/irzfub0nash5s9?cid=462)

I did use votes and reviews, they are not completely uninteresting :)

**what’s an insight?**

An insight is a significant correlation or trend between any 2 variables not listed in the questions above.

Also a lack of a correlation.

It would be nice if you could tell a story behind it.

Eg. You see a correlation that x is correlated with y. Think about the concepts, studying movies and what makes them successful in genres and years.

So an insight is something you can compress into a sentence or 2 and back it up with graphs and explain it to someone who has some understanding of that area of movies. And they can connect and make a story out of it.

Or explain why it’s surprising, why it contradicts your intuition. 1st show what the data shows you and then explain why it makes sense.

For non-surprising insights … basically tell the story of what you saw.

It needs to be something that people can understand and relate to.

**Grading**

TA’s will ask: did the student understand the question? Do what was asked?

Deal with the complex issues? Or ignore them.

Expecting some reasonable amount of diligence and retrospective, given the tools I’m going over.

Use your basic imagination as a data science. I would like to see more than bare bare minumim without going deeper.

**Back up graphs with chart of figures/numbers**