Capstone Project: Analysis of Movies

Arti Annaswamy
SlideRule: Foundations of Data Science Workshop

October 2015

Background

BA Business Administration

Excel

Cal State Univ, Chancellor's Office

Relational DBs

Excel

HTML / CSS

Deutsche Bank

SQL

Excel

MS Finance

Excel

Warner Bros. Consumer Products

Excel

SQL

R

Python

(Codecademy)

Foundations of Data Science Workshop

Topics Considered

Music

Travel

Movies

Available datasets

1. Last.fm
2. Million song dataset

1. KNOEMA.com
2. World Travel & Tourism Council

1. MovieLens
2. Yahoo Movie Ratings
3. Jester
4. Cornell dataset

Analysis ideas

Predict next song?
Rating? Skip Y/N?

Good for visualization, but what's a good statistical problem?

Predict movie rating (Netflix problem!)
Explore movie popularity vs time elapsed?
Predict Box Office?

Available fields

movies movieId, title, genres
ratings userId, movieId, rating, timestamp
links movieId, imdbId, tmdbId
tags movieId, tag, userId, timestamp

movielens
http://grouplens.org/datasets/movielens/

OMDb API
The Open Movie Database

IMDb: Title, Year, Rated, Released, Runtime, Genre, Director, Writer, Actors, Plot, Language, Country

Rotten Tomatoes: Fresh/Rotten rating, Critic rating, User Rating

YouTube Data API

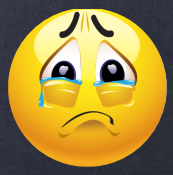
Search API
"movie+title+year+trailer"
Results: videoId, videoName

Get API
Results: videoId, Views, Likes, Dislikes, Favorites

IMDb

Interfaces data dump from IMDb:
11.41gb download, Python Anaconda setup, IMDbPy script, SQLObject setup, SQLite database and 54 hrs later -

Data scraper in R:
Results: imdbId, openingWeekend, Gross, Budget

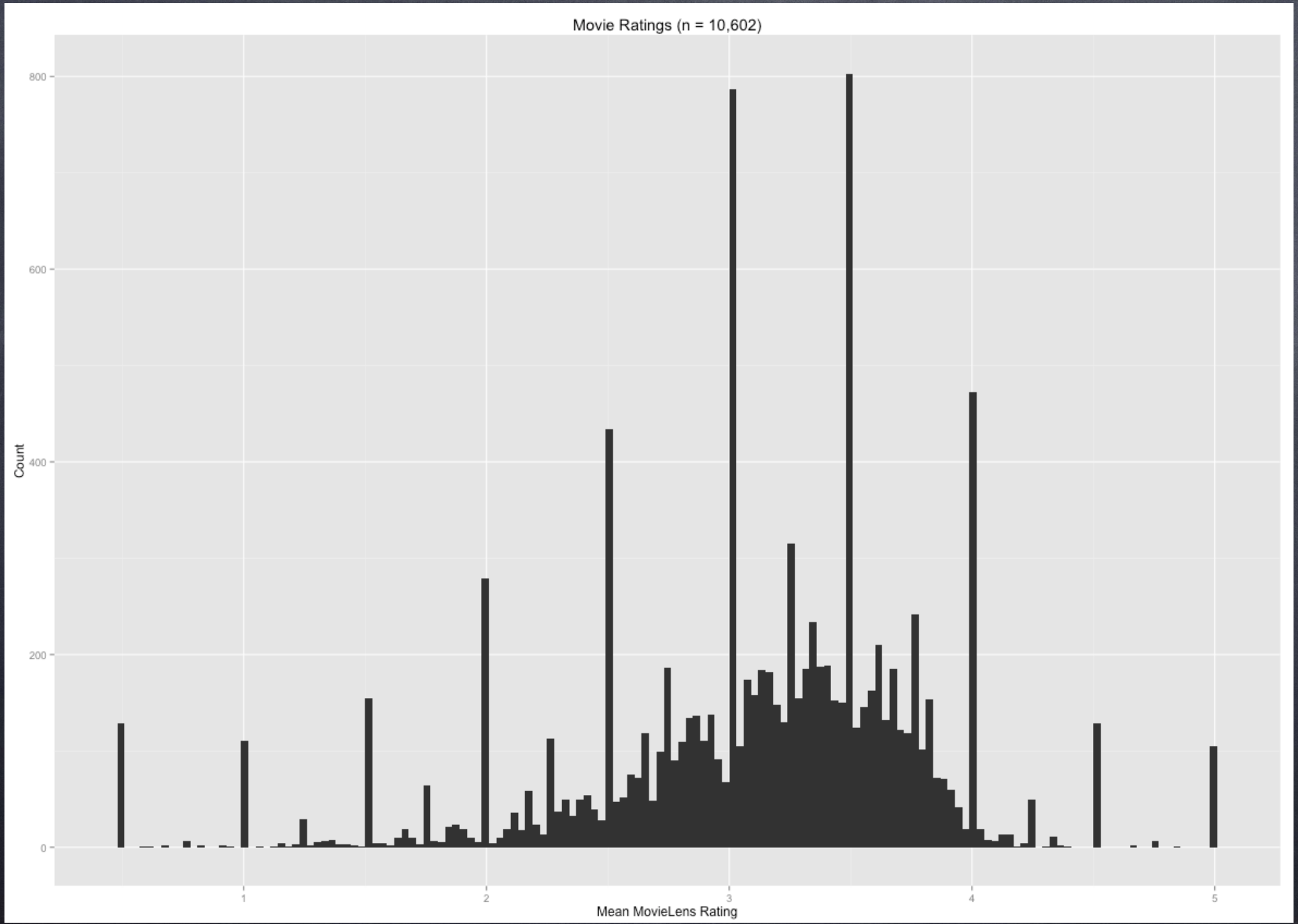


Final dataset

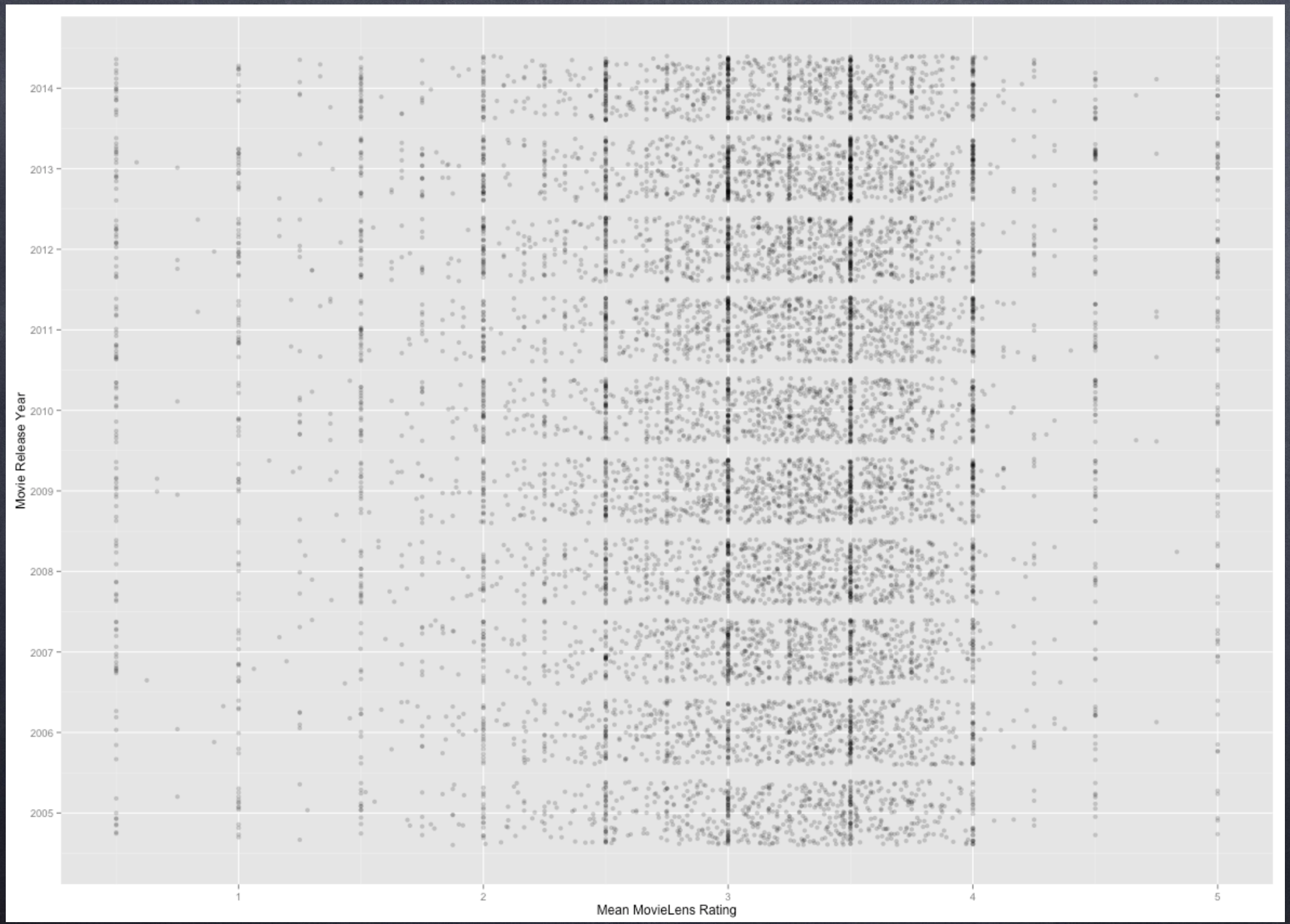
Column Name	Example Data
Year	2012
imdbId	tt1345836
opening_weekend	160887295
budget_est	250000000
gross	448130642
title	The Dark Knight Rises
Rated	PG-13
Released	7/20/12
Director	Christopher Nolan
Metascore	78
imdbRating	8.5
tomatoMeter	87
tomatoRating	8
tomatoUserRating	4.3
movieId	91529
searchString	the+dark+knight+rises+2012+trailer
ratingMean	3.995676293
nRat	6129
ratingMedian	4

Column Name	Example Data
viewCount	66211508
likeCount	197311
dislikeCount	6112
favCount	0
commentCount	150827
Actors	Christian Bale, Gary Oldman, Tom Hardy, Joseph Gordon-Levitt
Plot	Eight years after the Joker's reign of anarchy, the Dark Knight is forced to return from his imposed exile to save Gotham City from the brutal guerrilla terrorist Bane with the help of the enigmatic Catwoman.
tomatoConsensus	The Dark Knight Rises is an ambitious, thoughtful, and potent action film that concludes Christopher Nolan's franchise in spectacular fashion.
genre columns	1
runtime	165
rated columns	0
ratedTVMA	0
grp	4
Inflation.Factor	2.439129696
viewCount_adj	27145546.26

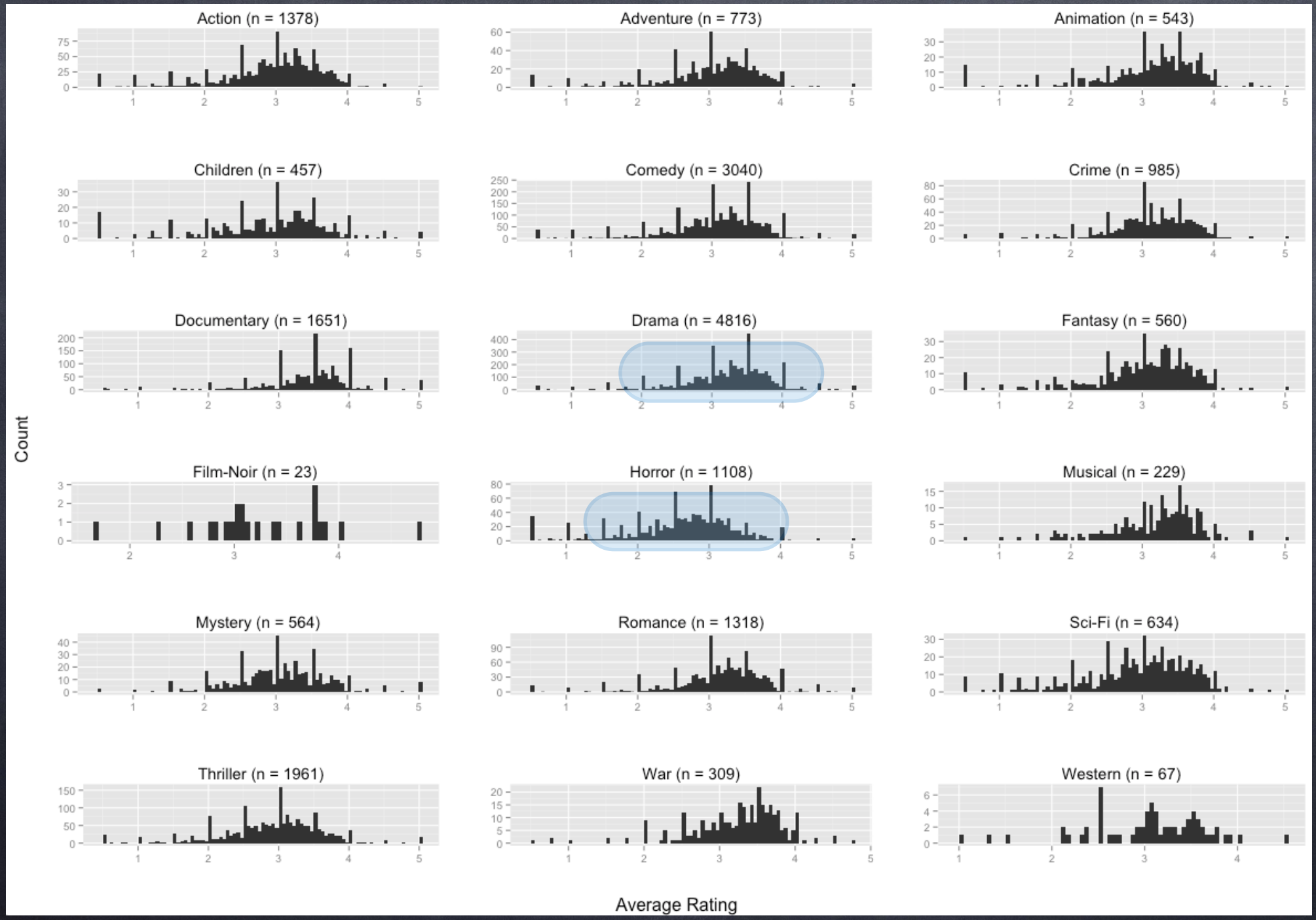
Average MovieLens Rating Distribution



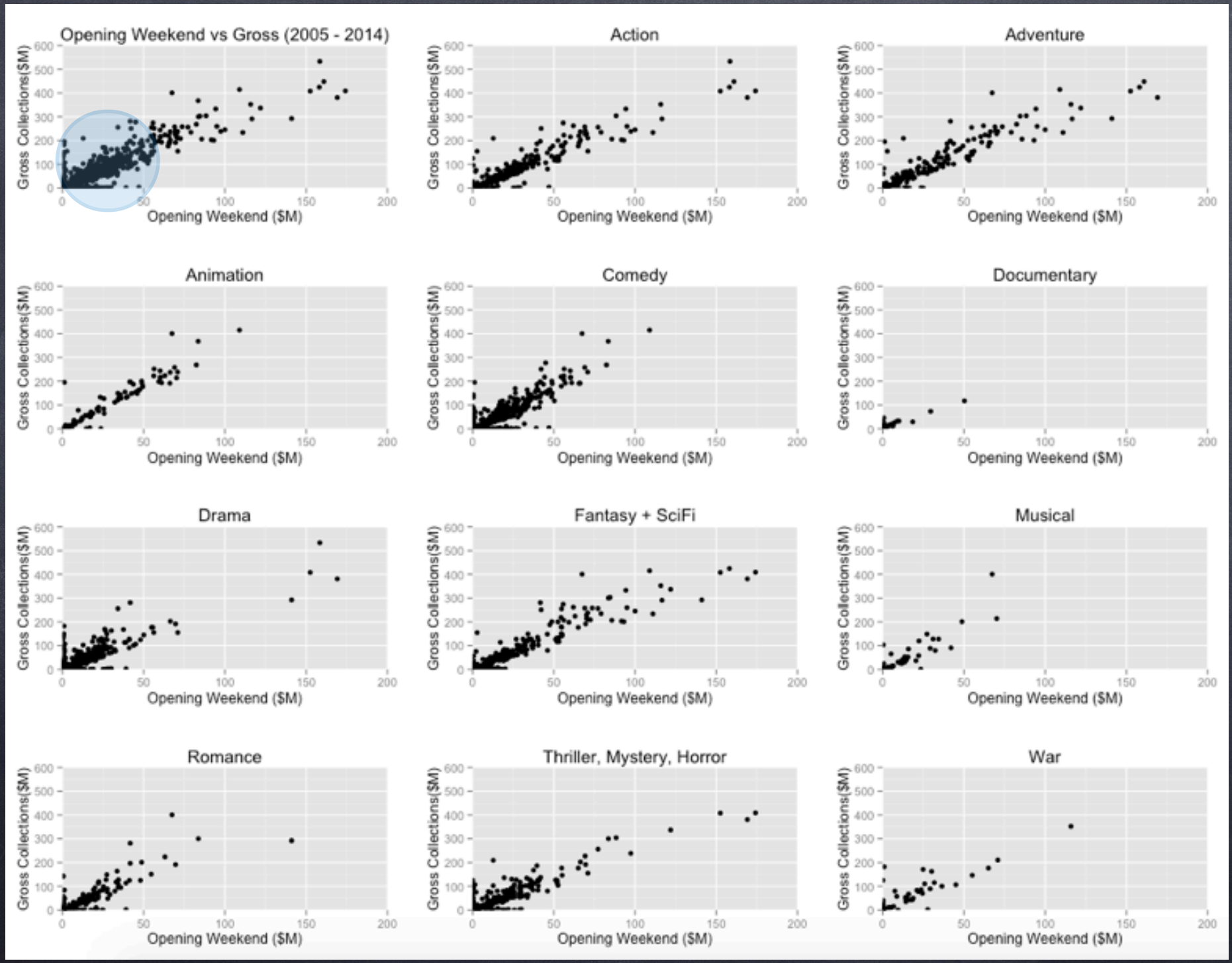
Average MovieLens Rating Distribution



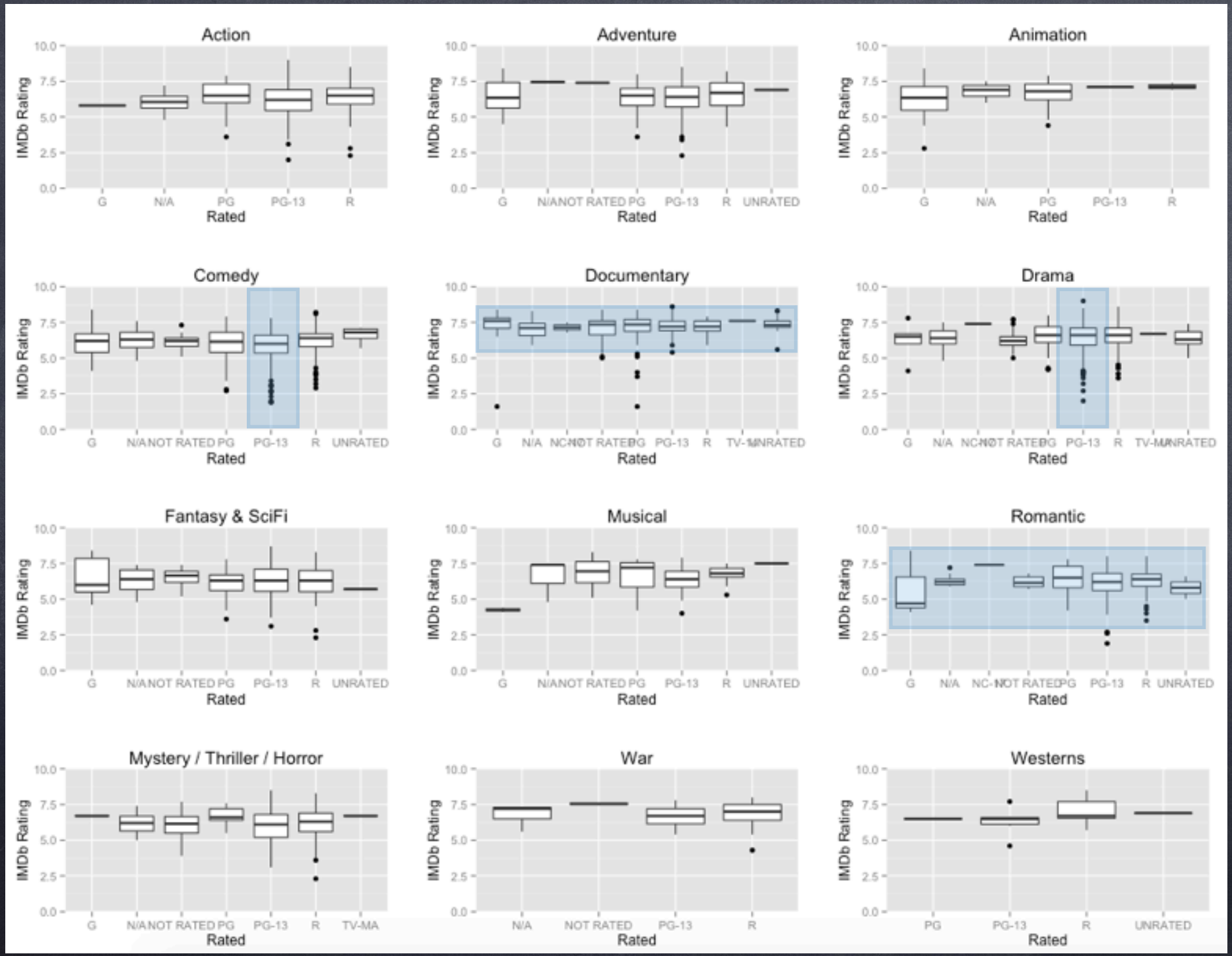
Average MovieLens Rating Distribution



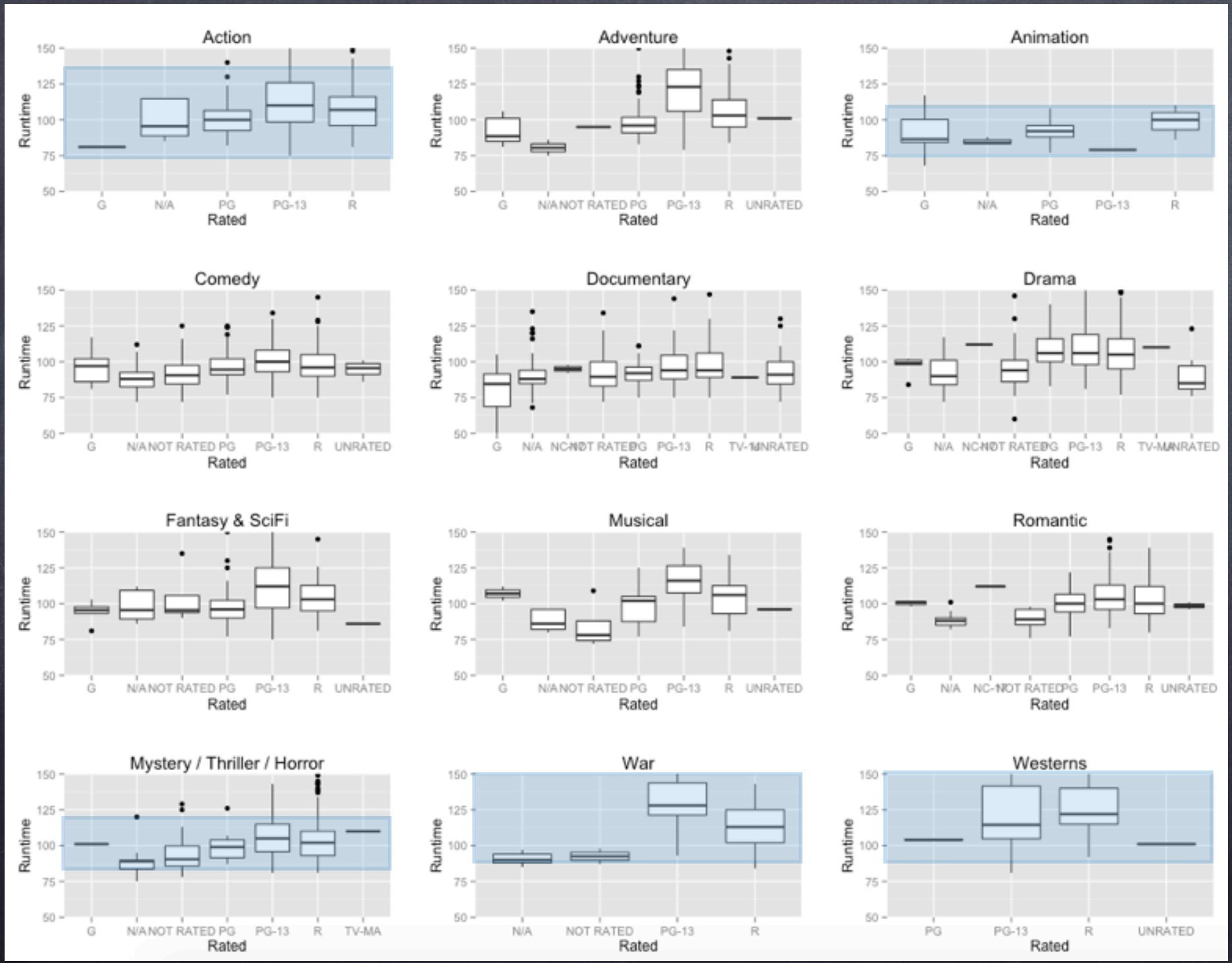
Gross Collections and Opening Weekend, by Genre



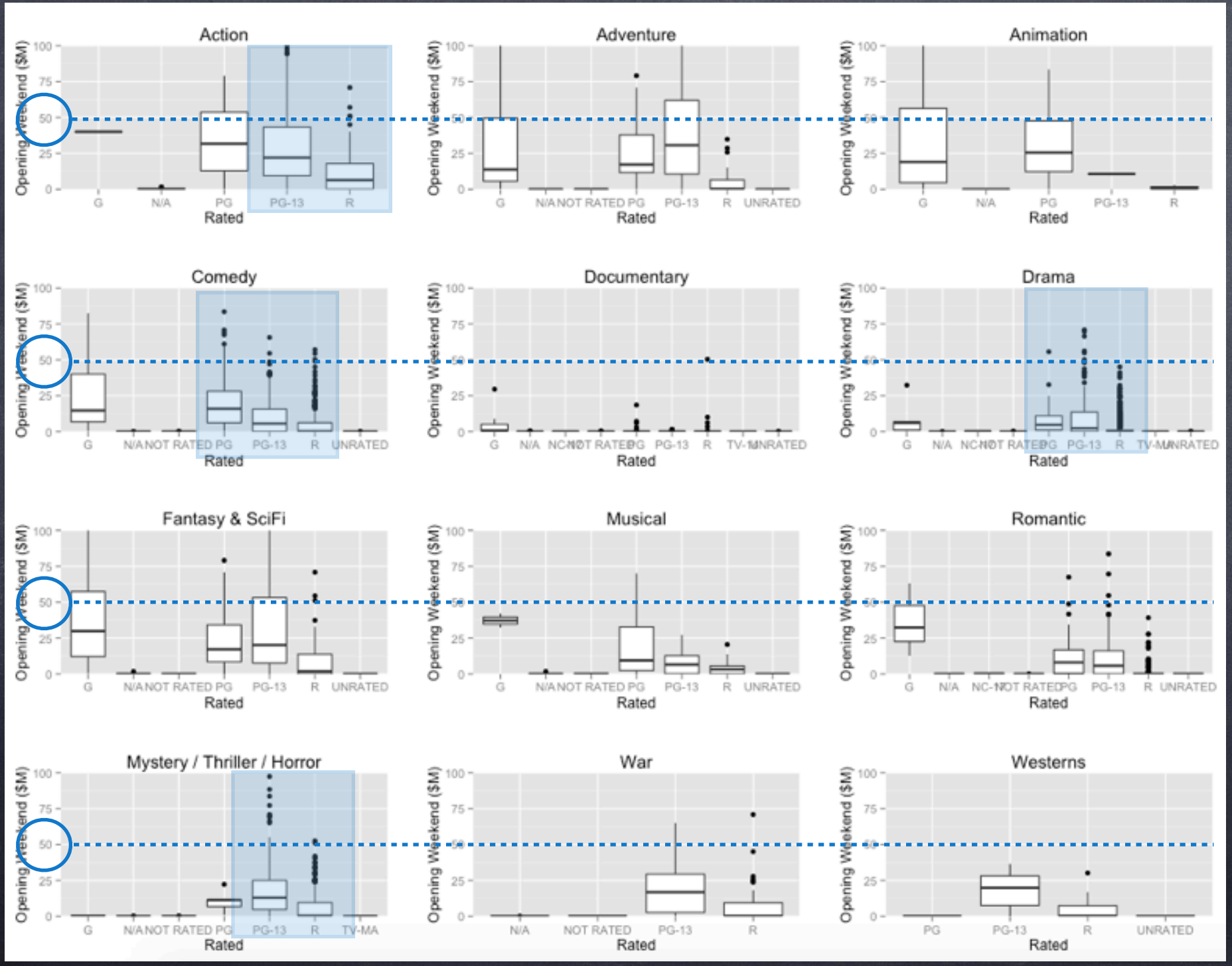
IMDb User Rating by MPAA Rating & Genre



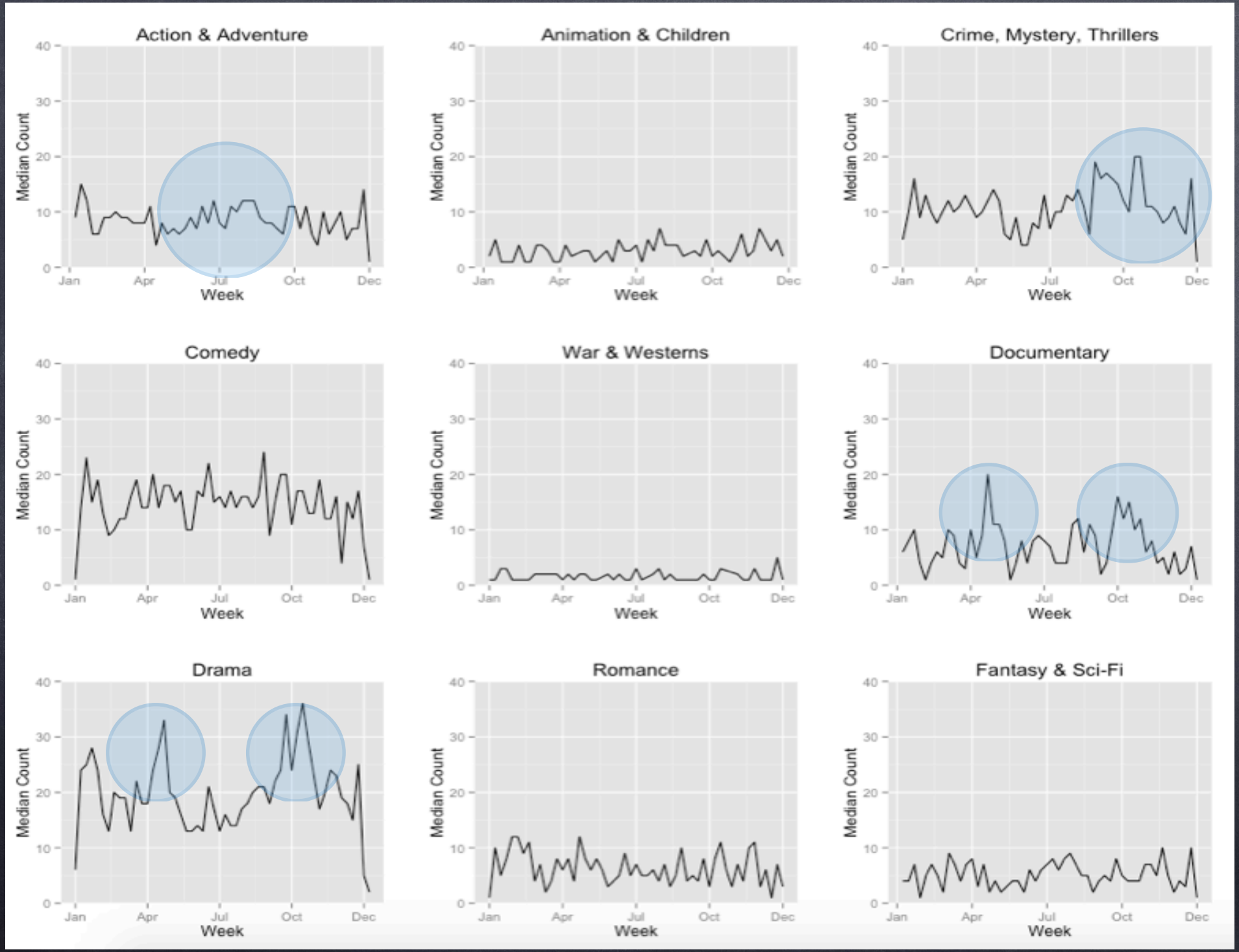
Movie Runtime (Minutes) by MPAA Rating & Genre



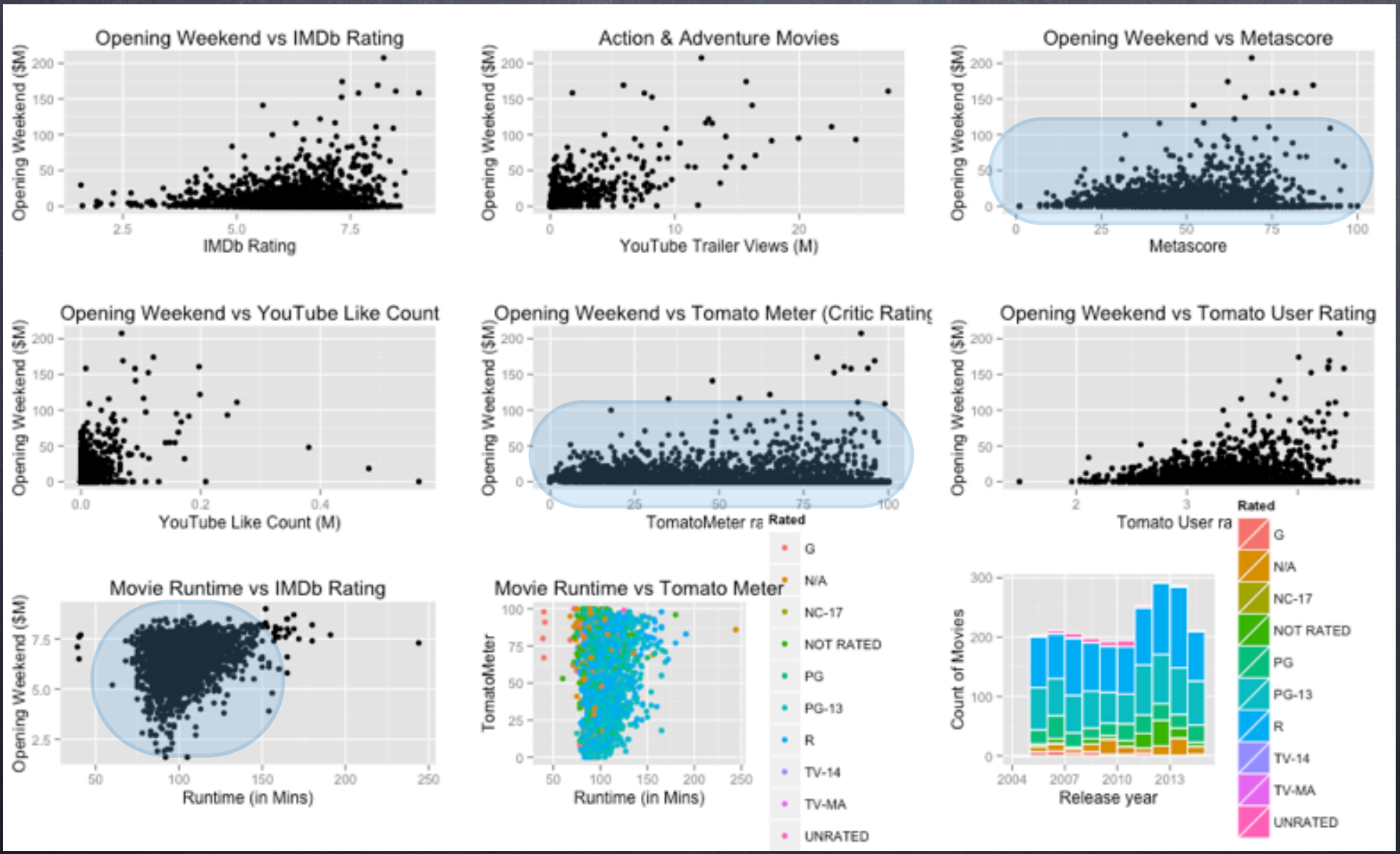
Opening Weekend ($M) by MPAA Rating & Genre



Median Count of Movies Released, by Month



Other visualizations…



Analysis Methods

- Simple Linear Regression
- Stepwise Regression using step()
- Text Analysis "Bag of Words"
  - Tomato Consensus - summary of Critics Reviews on Rotten Tomatoes
  - Plot Summary
  - Actors (top 5 credited) & Directors
- k-Means Clustering

Text Analysis - Tomato Consensus

varNames	Pr(>|t|)
compel	0.157
smart	0.153
suffer	0.147
explor	0.128
laugh	0.109
fun	0.082
talent	0.057
effect	0.051
ambiti	0.027
origin	0.020
director	0.017
man	0.012
comic	0.011
success	0.009
part	0.005
thrill	0.005
world	0.004
surpris	0.000
impress	0.000
entertain	0.000
franchis	0.000

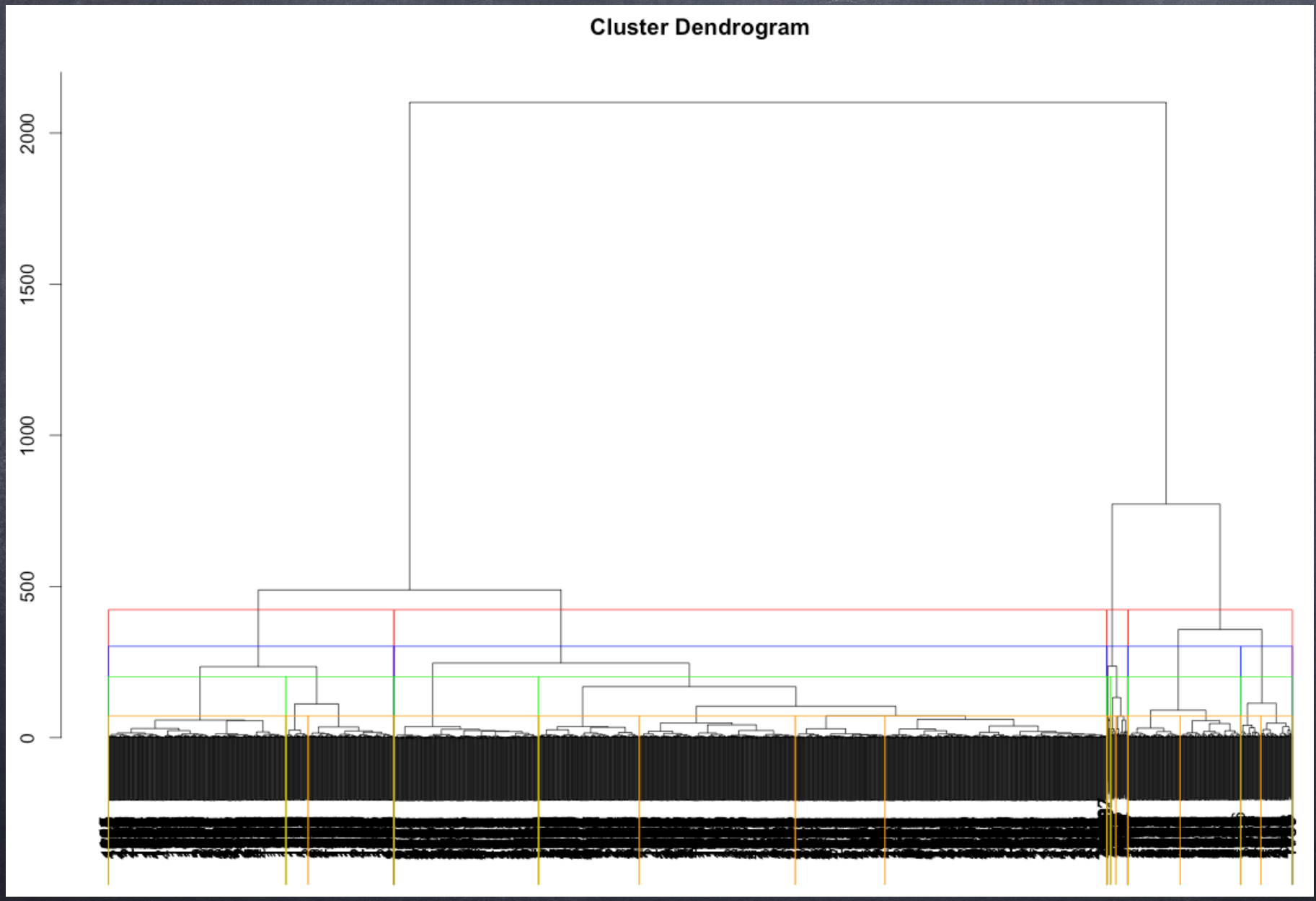
Text Analysis - Plot Summary

varNames	Pr(>|t|)
start	0.267
son	0.249
goe	0.160
stori	0.147
young	0.145
follow	0.106
come	0.101
war	0.008
take	0.002
fight	0.000

Text Analysis - Cast & Crew

varNames	Pr(>|t|)
jasonstatham	0.214
stevensoderbergh	0.183
benaffleck	0.168
guypearce	0.166
ratingMean	0.162
stevebuscemi	0.121
camerondiaz	0.094
shialabeouf	0.076
stevecarell	0.064
zacefron	0.062
georgeclooney	0.059
helenmirren	0.041
eddiemurphy	0.026
jenniferlawrence	0.026
kevinjames	0.011
benstiller	0.007
tylerperry	0.007
willferrell	0.004
kristenstewart	0.002
willarnett	0.000
keanureeves	0.000
sandrabullock	0.000
robertdowneyjr	0.000

Clustering



Analyses Results

Test Details	# of Var	Train R2	Test R2	fStat
Simple Linear Regression on full dataset	37	0.581	0	75.0156 on 42 and 2203 DF
Linear Regression on full dataset, with stepwise regression	19	0.581	0	174.1394 on 18 and 2227
Linear Regression on training set #1, with stepwise regression	24	0.597	0.505	102.1876 on 23 and 1548
Linear Regression on training set #2, with stepwise regression	24	0.592	0.559	99.9151 on 23 and 1548 DF
Tomato Consensus - Bag of Words, LM, on Training Set #1	204	0.640	0.469	14.7777 on 203 and 1368
Tomato Consensus - Bag of Words, LM, on Training Set #1, less variables	95	0.658	0.559	33.2125 on 94 and 1477 DF
Tomato Consensus - Bag of Words, LM, on Training Set #1, with stepwise	72	0.657	0.559	43.4175 on 71 and 1500 DF
Tomato Consensus - Bag of Words, LM, on Training Set #2	72	0.635	0	39.4788 on 71 and 1500 DF
Tomato Consensus - Bag of Words, LM, on Training Set #2 with stepwise	49	0.636	0.508	58.2906 on 48 and 1523 DF
Plot Summary - Bag of Words, LM, on Training Set #1	170	0.598	0.485	14.8183 on 169 and 1402
Plot Summary - Bag of Words, LM, on Training Set #1 with stepwise regr	50	0.614	0.485	52.0749 on 49 and 1522 DF
Plot Summary - Bag of Words, LM, on Training Set #2 with stepwise regr	35	0.587	0.586	66.7652 on 34 and 1537 DF
Cast & Director - Bag of Words, LM, on Training Set#1	199	0.646	0.457	15.459 on 198 and 1373 DF
Cast & Director - Bag of Words, LM, on Training Set#1 with stepwise regr	73	0.662	0.457	43.7835 on 72 and 1499 DF
Cast & Director - Bag of Words, LM, on Training Set#2	73	0.639	0	39.5807 on 72 and 1499 DF
Cast & Director - Bag of Words, LM, on Training Set#2 with stepwise regr	55	0.640	0.551	52.7757 on 54 and 1517 DF
Clustering, LM, on Training Set#1 - Group 1	47	0.614	0.253	13.9661 on 46 and 329 DF
Clustering, LM, on Training Set#1 - Group 2	37	0.388	0.245	17.5478 on 36 and 905 DF
Clustering, LM, on Training Set#1 - Group 3	45	0.694	0.130	12.6023 on 44 and 181 DF
Clustering, LM, on Training Set#1 - Group 4	118	NaN	0	NaN on 27 and 0 DF


FROM THE DIRECTOR OF GLADIATOR
EXODUS
GODS AND KINGS
DECEMBER 12
FROM THE CREATORS OF
INSIDE OUT
MEET THE LITTLE VOICES INSIDE YOUR HEAD.
JUNE 19
DWAYNE JOHNSON
SAN ANDREAS
MAY 29
JURASSIC WORLD
JUNE 12
LLEGA EL MOMENTO DE LA VENGANZA
FAST & FURIOUS 7
2 DE ABRIL EN CINES
HOW TO TRAIN YOUR DRAGON 2
BIG HERO 6
GO ROGUE
MISSION: IMPOSSIBLE
ROGUE NATION
JULY 31
MMXL
CINDERELLA
MARCH 13
LIVE. DIE. REPEAT
THE AMAZING SPIDER-MAN 2
IN 3D
CAPTAIN AMERICA
THE WINTER SOLDIER
2014
ANT-MAN
EX_MACHINA
APRIL 24
QUANTUM OF SOLACE
MAX
FURY ROAD
2015
INTO THE WOODS
BE CAREFUL WHAT YOU WISH FOR
CHRISTMAS
THE LEGO MOVIE
ASSEMBLING IN 2014
X-MEN
DAYS OF FUTURE PAST
THE HOBBIT
FANTASTIC 4
08.07.15
TOMORROWLAND
IRON MAN
AVENGERS
AGE OF ULTRON
COMING SOON
POMPEII
NO WARNING
NO ESCAPE
THE MAZE RUNNER
BASED ON THE BEST-SELLING NOVEL
SEPTEMBER 19
INTERSTELLAR
COMING SOON
GUARDIANS
OF THE
GALAXY

thank you!