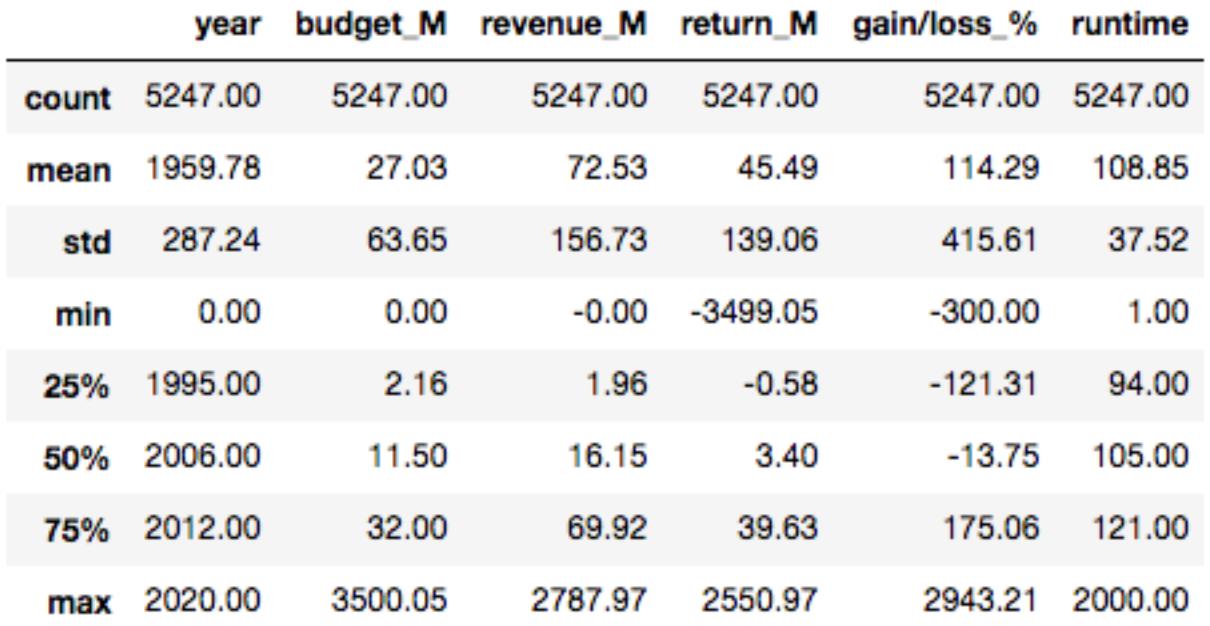
What makes movie more profitable?

Group project: Mulbah, Zach, Evgeniya

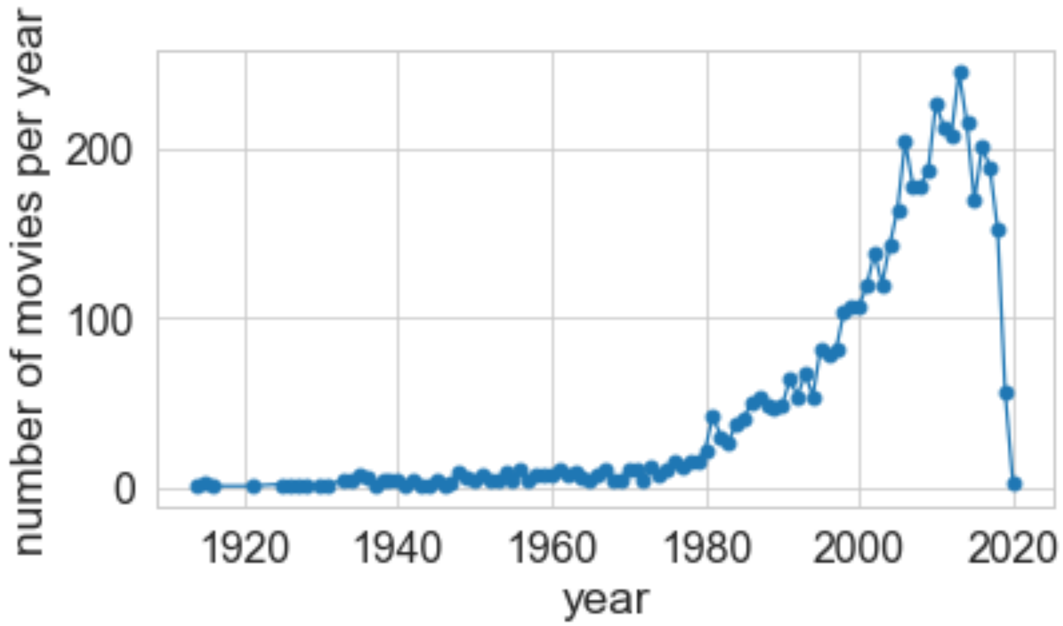
Getting the data

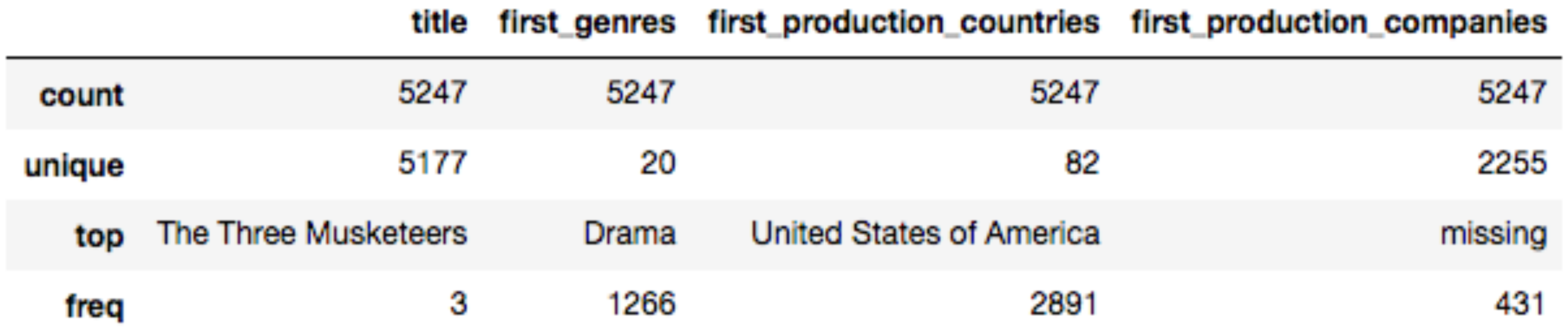
- Evgeniya: https://www.themoviedb.org (via API)
- Mulbah: http://www.the-numbers.com (copy to excel) and https://www.themoviedb.org (via API)
- Zach: http://www.the-numbers.com (web scraping)

Analysis of themoviedb data



	year	budget_M	revenue_M	return_M	gain/loss_%	runtime
count	5247.00	5247.00	5247.00	5247.00	5247.00	5247.00
mean	1959.78	27.03	72.53	45.49	114.29	108.85
std	287.24	63.65	156.73	139.06	415.61	37.52
min	0.00	0.00	-0.00	-3499.05	-300.00	1.00
25%	1995.00	2.16	1.96	-0.58	-121.31	94.00
50%	2006.00	11.50	16.15	3.40	-13.75	105.00
75%	2012.00	32.00	69.92	39.63	175.06	121.00
max	2020.00	3500.05	2787.97	2550.97	2943.21	2000.00



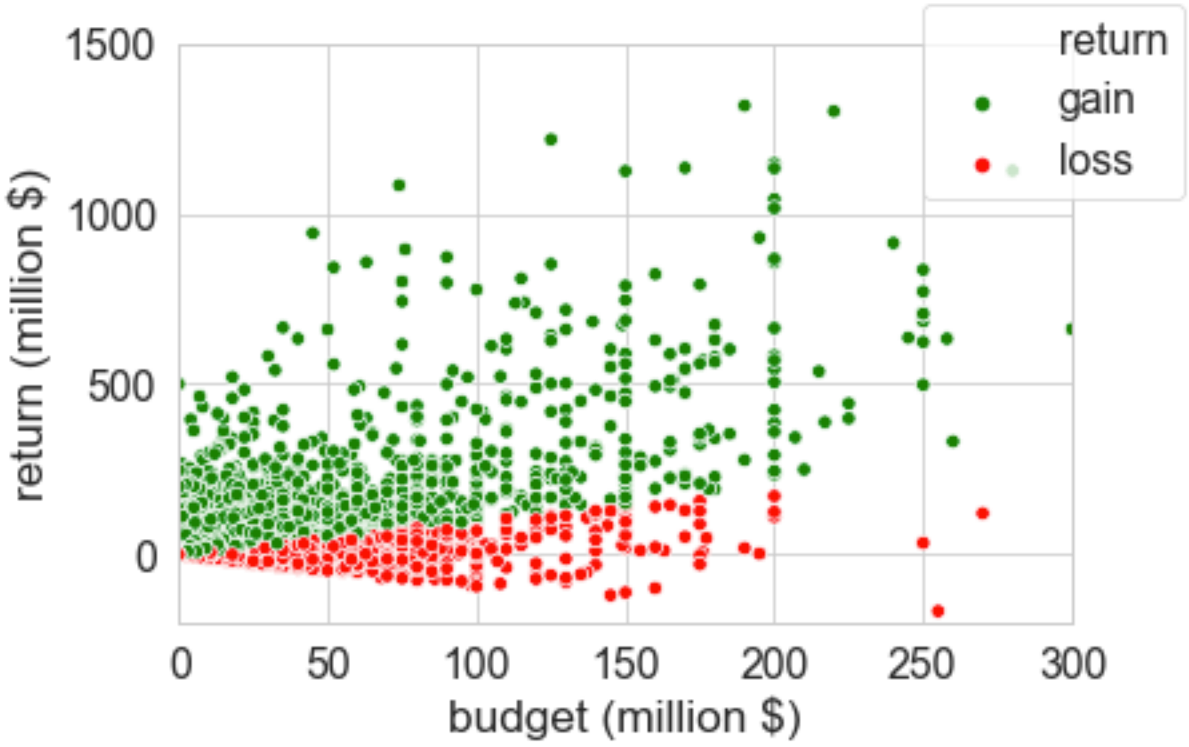


	title	first_genres	first_production_countries	first_production_companies
count	5247	5247	5247	5247
unique	5177	20	82	2255
top	The Three Musketeers	Drama	United States of America	missing
freq	3	1266	2891	431

Questions

- How budget influence the return?
- What is the influence of the production country?
- What is the influence of the production company?
- Select: budget category, company, genre.

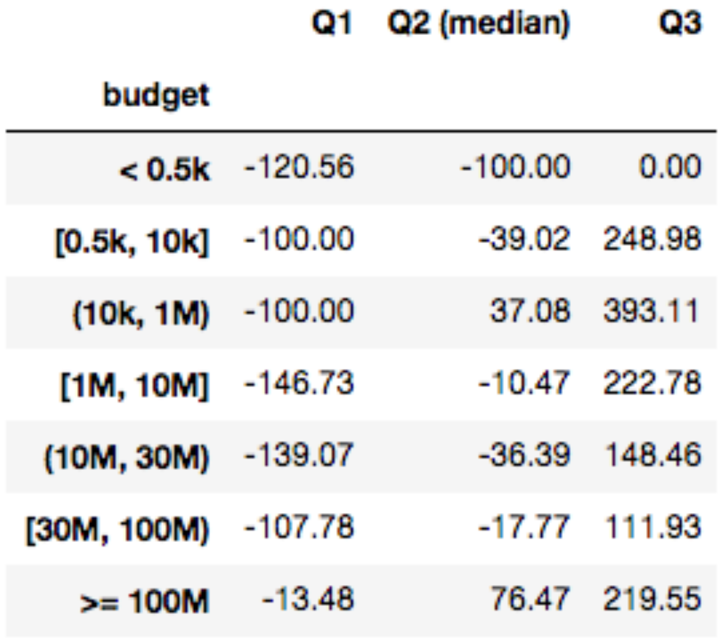
Return versus budget



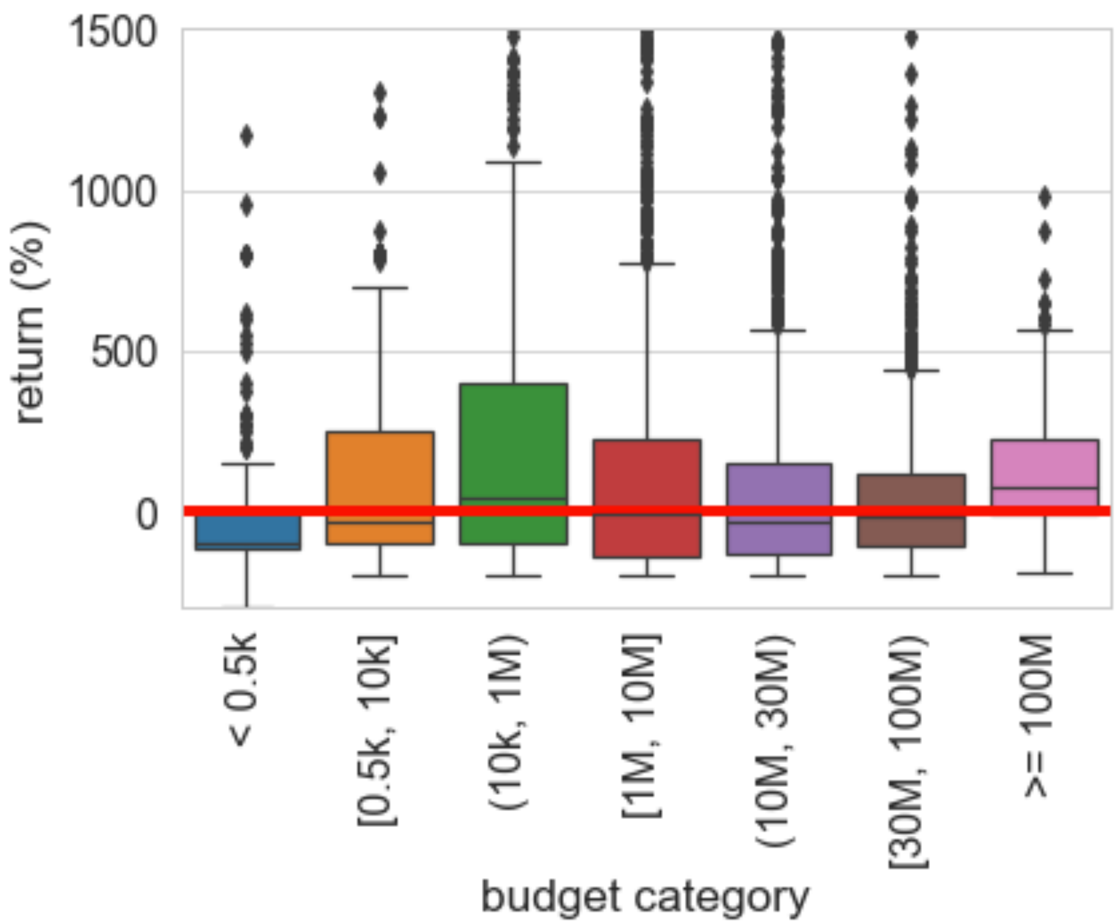
return = (revenue - budget)/budget

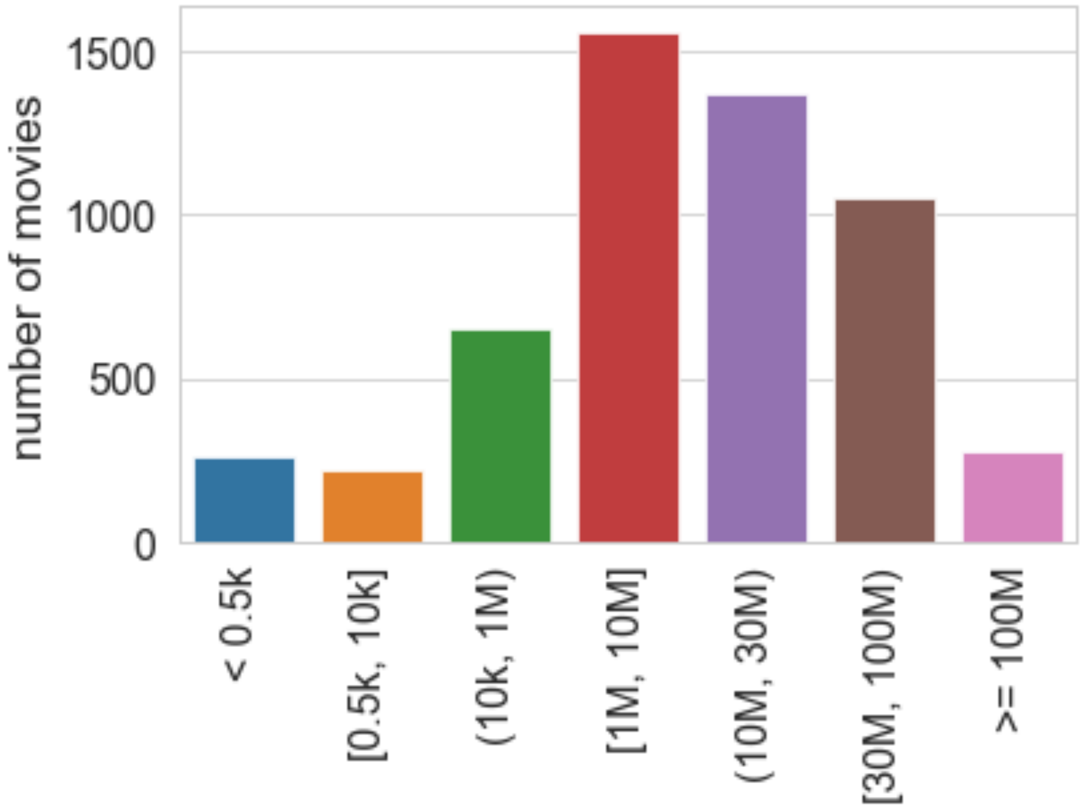
50.7% of movies have negative return.

Return versus budget

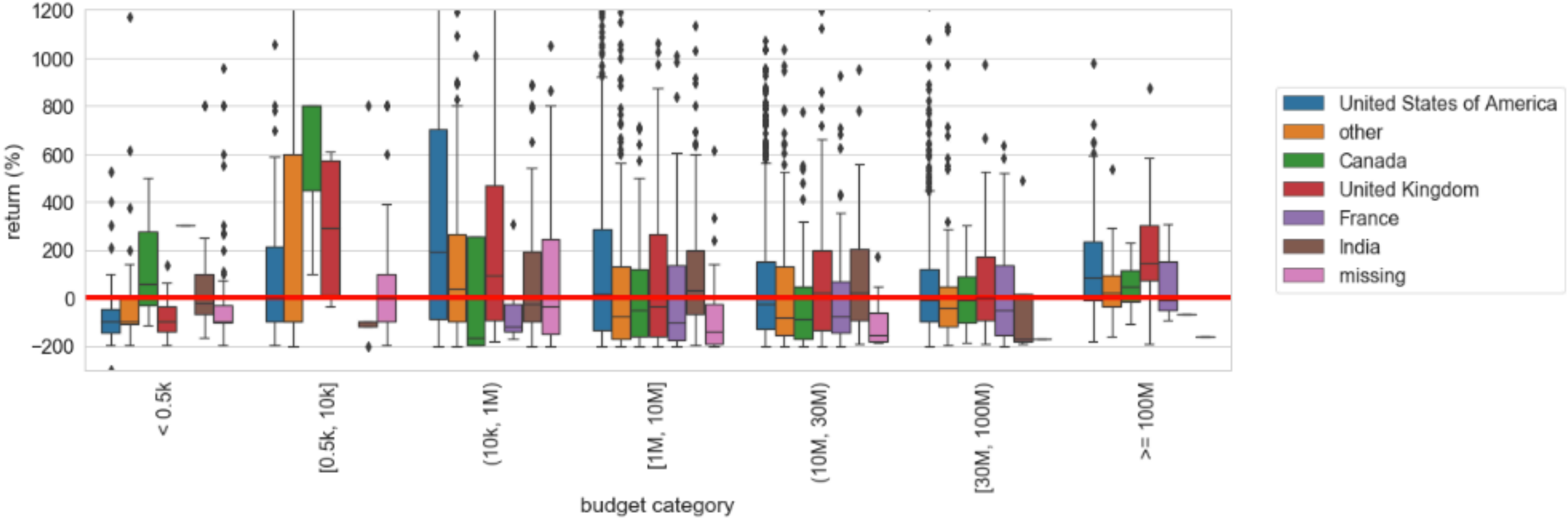


budget	Q1	Q2 (median)	Q3
< 0.5k	-120.56	-100.00	0.00
[0.5k, 10k]	-100.00	-39.02	248.98
(10k, 1M)	-100.00	37.08	393.11
[1M, 10M]	-146.73	-10.47	222.78
(10M, 30M)	-139.07	-36.39	148.46
[30M, 100M)	-107.78	-17.77	111.93
>= 100M	-13.48	76.47	219.55

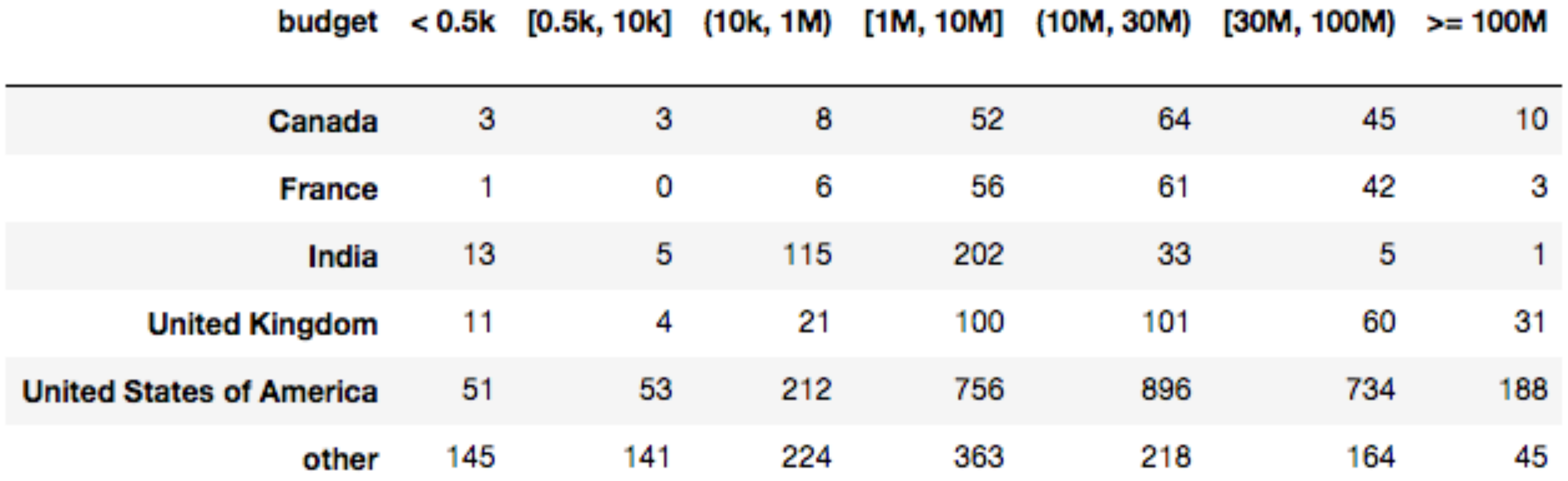




Return versus budget for top 5 countries

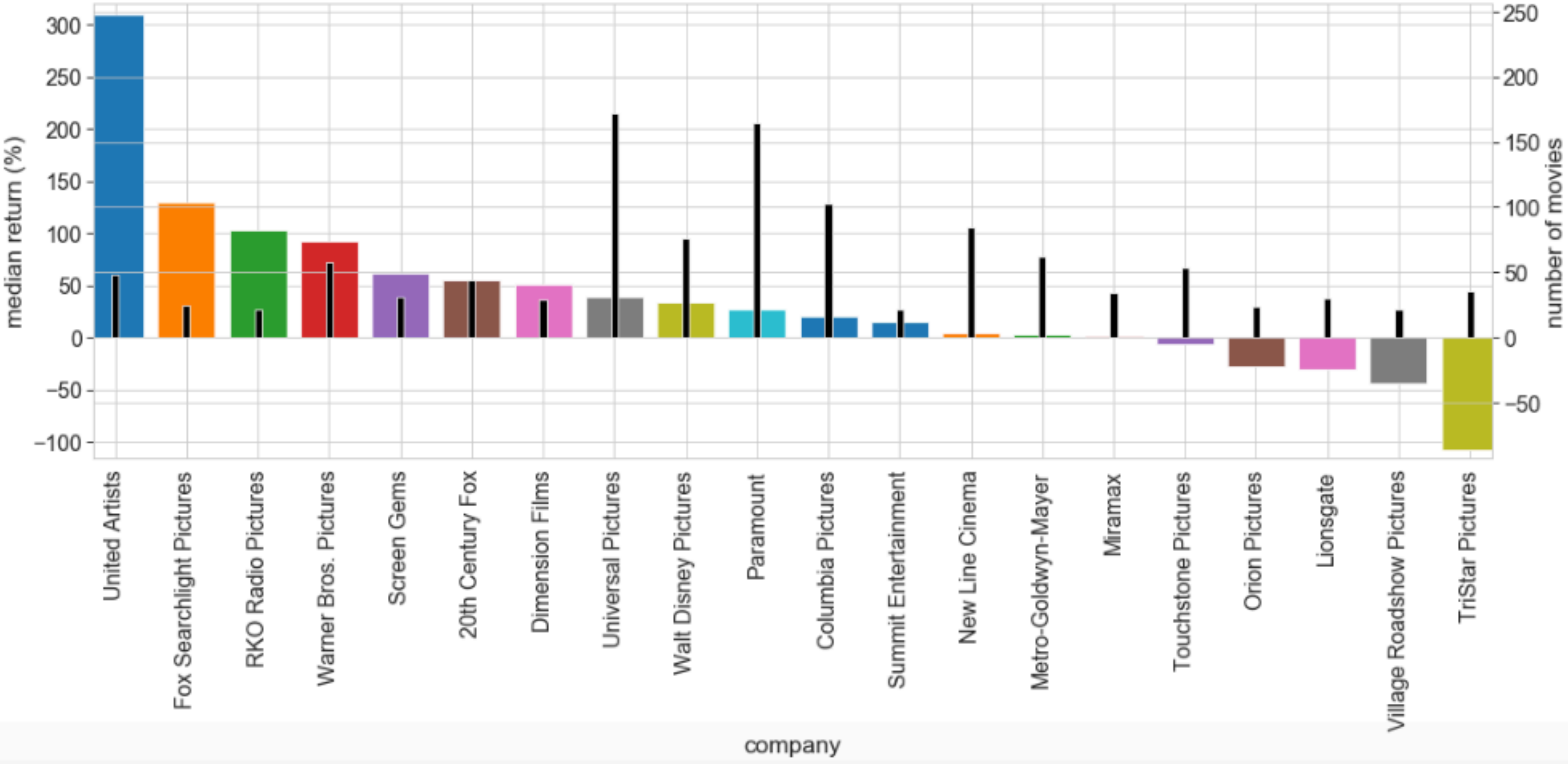


number of movies in a given category

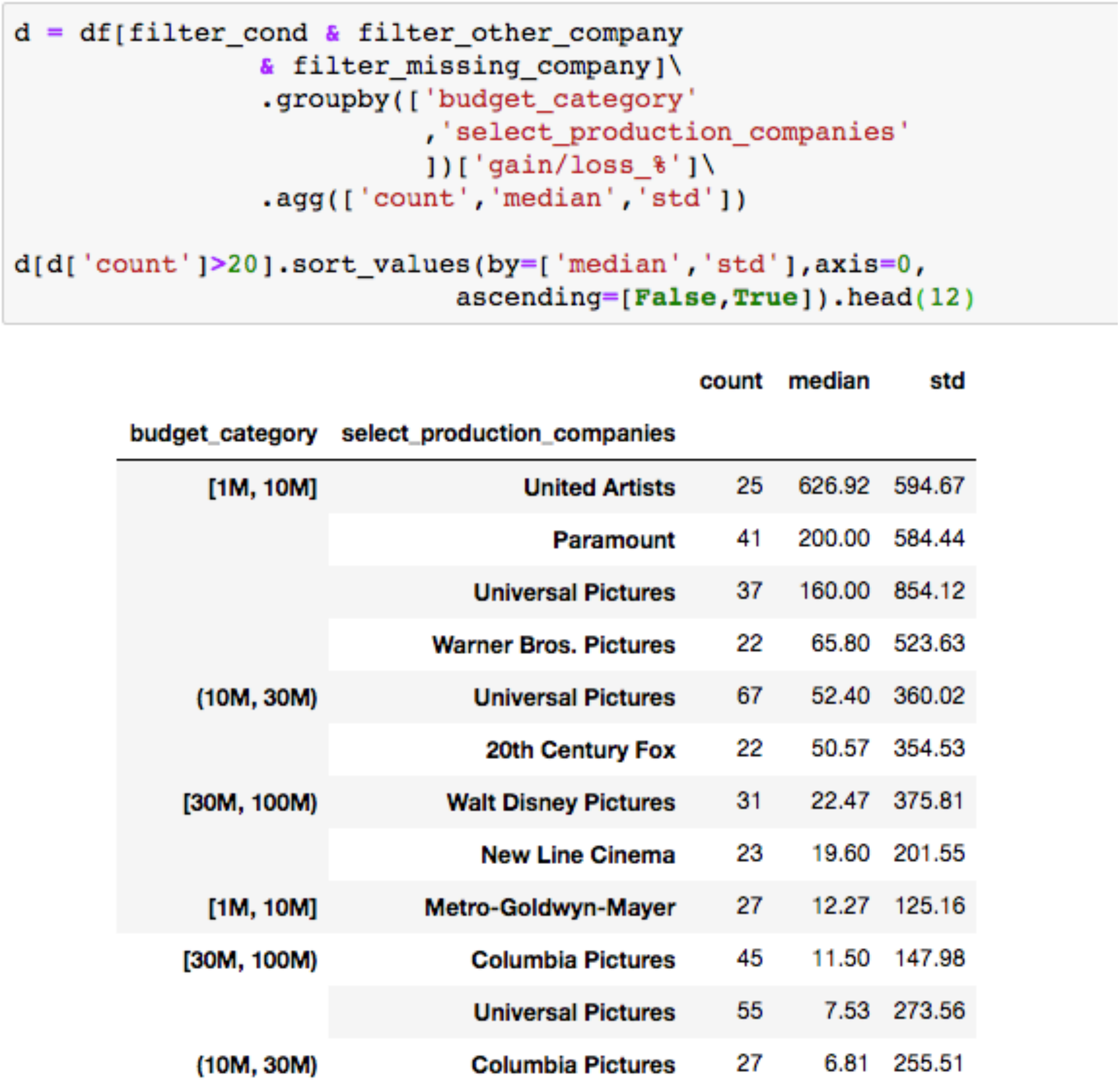


budget	< 0.5k	[0.5k, 10k]	(10k, 1M)	[1M, 10M]	(10M, 30M)	[30M, 100M)	>= 100M
Canada	3	3	8	52	64	45	10
France	1	0	6	56	61	42	3
India	13	5	115	202	33	5	1
United Kingdom	11	4	21	100	101	60	31
United States of America	51	53	212	756	896	734	188
other	145	141	224	363	218	164	45

Return versus budget for top 20 companies



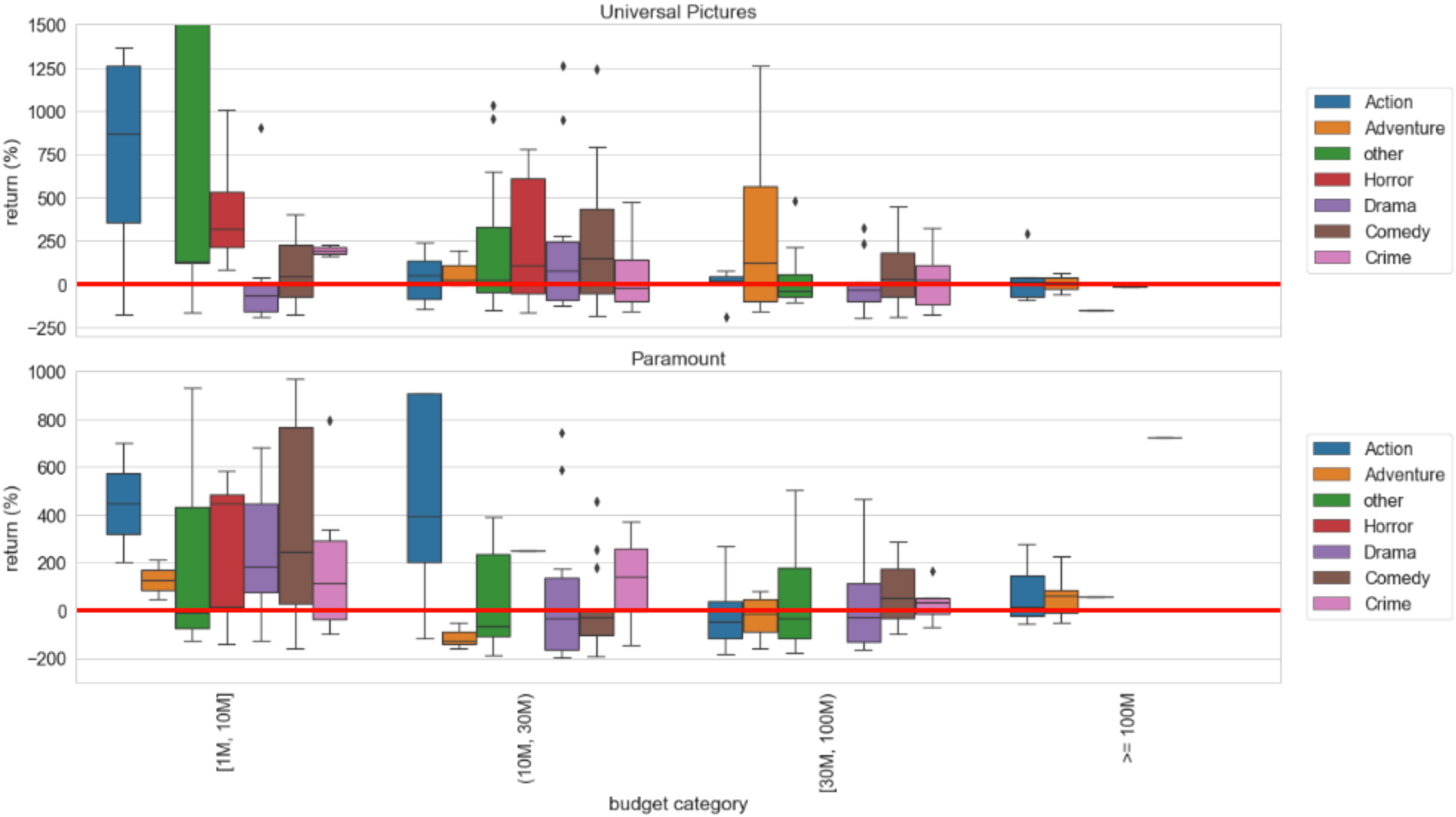
Select budget category and company



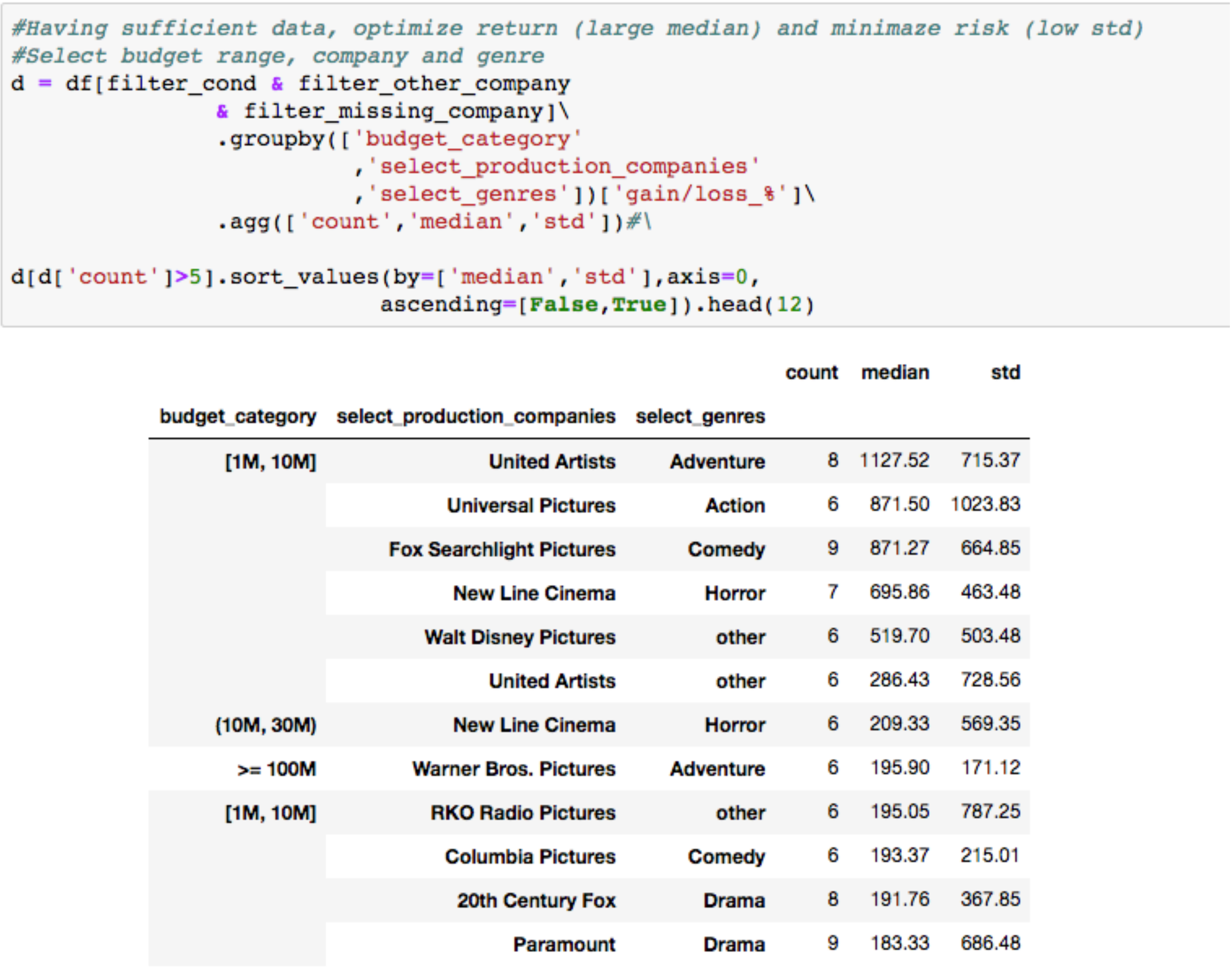
```
d = df[filter_cond & filter_other_company
               & filter_missing_company]\
               .groupby(['budget_category'
                         ,'select_production_companies'
                         ])['gain/loss_%']\
               .agg(['count','median','std'])

d[d['count']>20].sort_values(by=['median','std'],axis=0,
                         ascending=[False,True]).head(12)
```

		count	median	std
budget_category	select_production_companies			
[1M, 10M]	United Artists	25	626.92	594.67
	Paramount	41	200.00	584.44
	Universal Pictures	37	160.00	854.12
	Warner Bros. Pictures	22	65.80	523.63
(10M, 30M)	Universal Pictures	67	52.40	360.02
	20th Century Fox	22	50.57	354.53
[30M, 100M)	Walt Disney Pictures	31	22.47	375.81
	New Line Cinema	23	19.60	201.55
[1M, 10M]	Metro-Goldwyn-Mayer	27	12.27	125.16
[30M, 100M)	Columbia Pictures	45	11.50	147.98
	Universal Pictures	55	7.53	273.56
(10M, 30M)	Columbia Pictures	27	6.81	255.51

Select budget category, company, genre



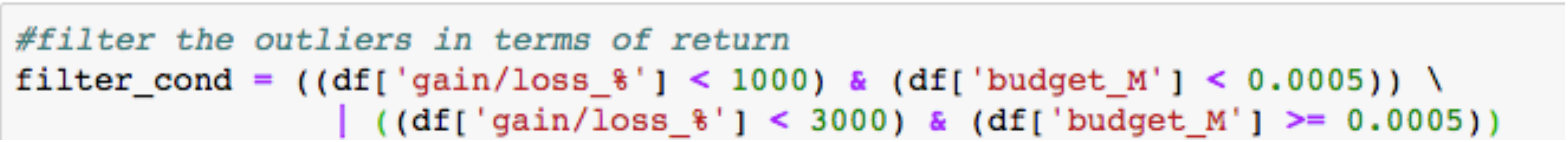
Select budget category, company, genre



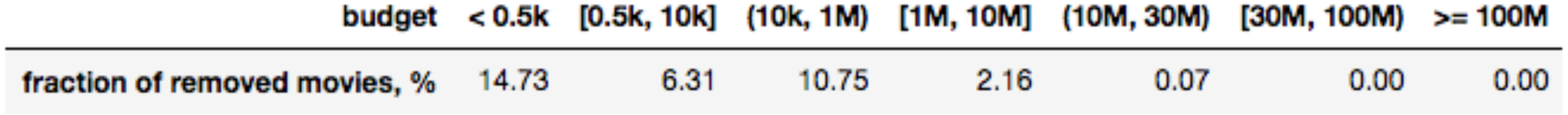
```
#Having sufficient data, optimize return (large median) and minimaze risk (low std)
#Select budget range, company and genre
d = df[filter_cond & filter_other_company
               & filter_missing_company]\
               .groupby(['budget_category'
                         ,'select_production_companies'
                         ,'select_genres'])['gain/loss_%']\
               .agg(['count','median','std'])#\

d[d['count']>5].sort_values(by=['median','std'],axis=0,
                           ascending=[False,True]).head(12)
```

			count	median	std
budget_category	select_production_companies	select_genres			
[1M, 10M]	United Artists	Adventure	8	1127.52	715.37
	Universal Pictures	Action	6	871.50	1023.83
	Fox Searchlight Pictures	Comedy	9	871.27	664.85
	New Line Cinema	Horror	7	695.86	463.48
	Walt Disney Pictures	other	6	519.70	503.48
	United Artists	other	6	286.43	728.56
(10M, 30M)	New Line Cinema	Horror	6	209.33	569.35
>= 100M	Warner Bros. Pictures	Adventure	6	195.90	171.12
[1M, 10M]	RKO Radio Pictures	other	6	195.05	787.25
	Columbia Pictures	Comedy	6	193.37	215.01
	20th Century Fox	Drama	8	191.76	367.85
	Paramount	Drama	9	183.33	686.48

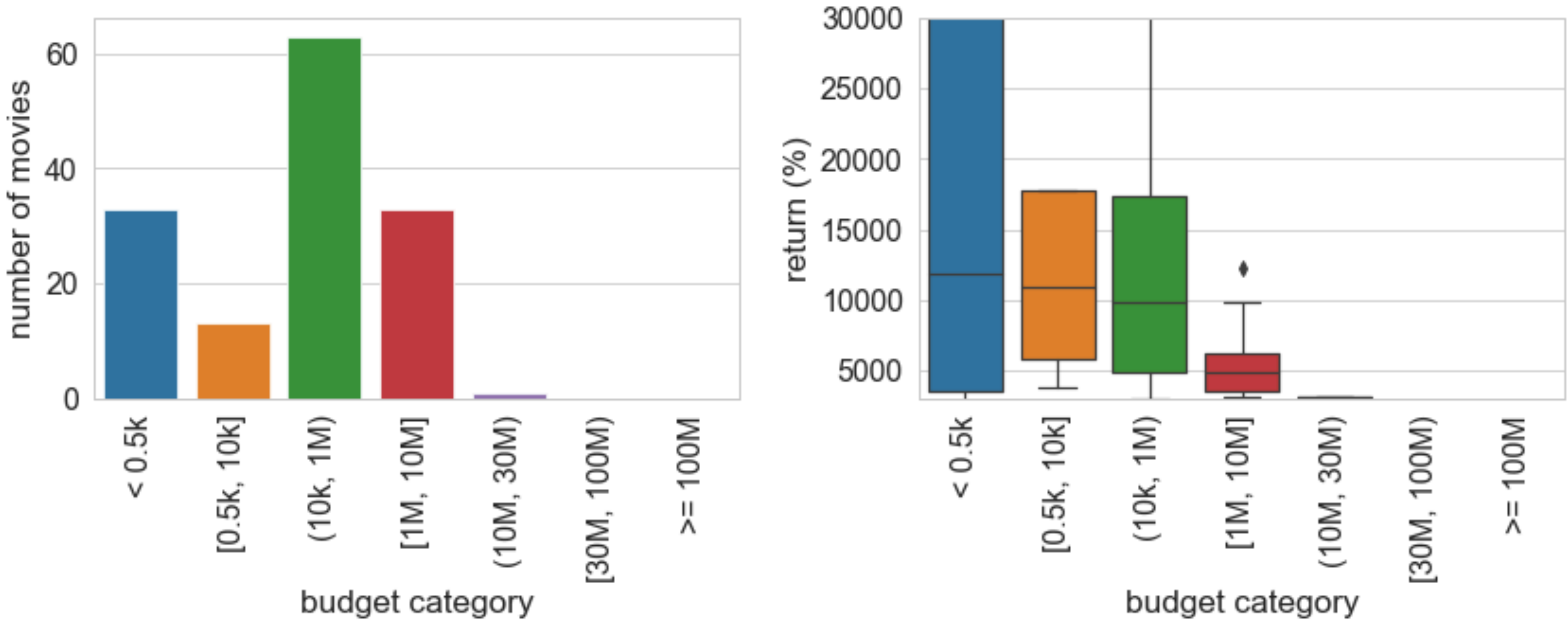
Extra: Filtering procedure

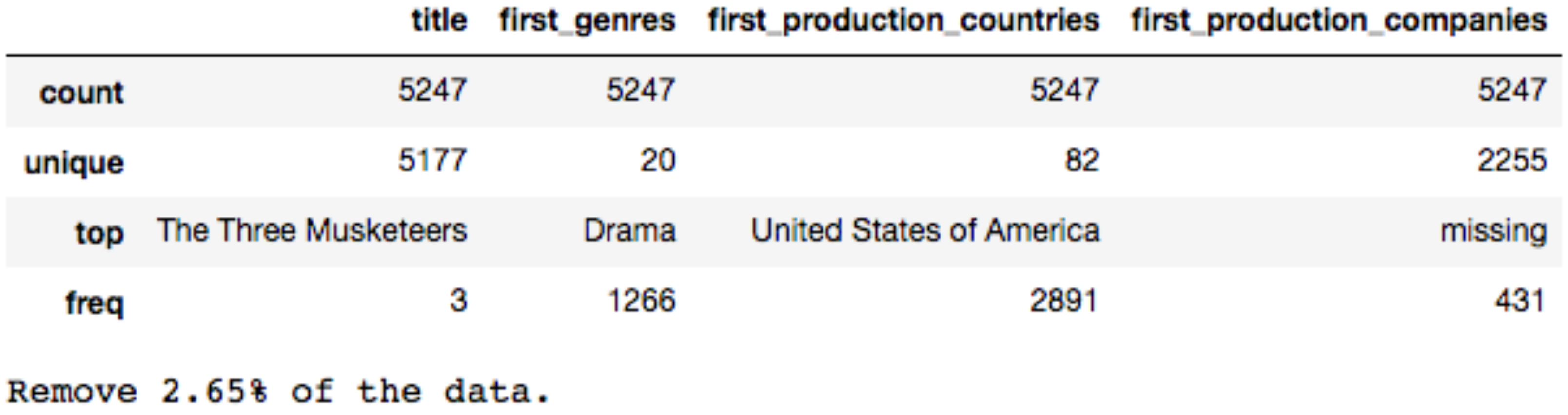


```
#filter the outliers in terms of return
filter_cond = ((df['gain/loss_%'] < 1000) & (df['budget_M'] < 0.0005)) \
                 | ((df['gain/loss_%'] < 3000) & (df['budget_M'] >= 0.0005))
```



budget	< 0.5k	[0.5k, 10k]	(10k, 1M)	[1M, 10M]	(10M, 30M)	[30M, 100M)	>= 100M
fraction of removed movies, %	14.73	6.31	10.75	2.16	0.07	0.00	0.00

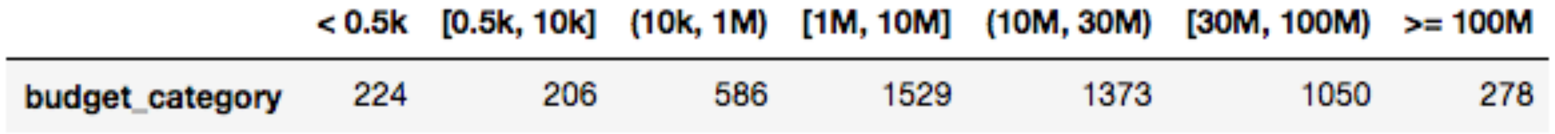




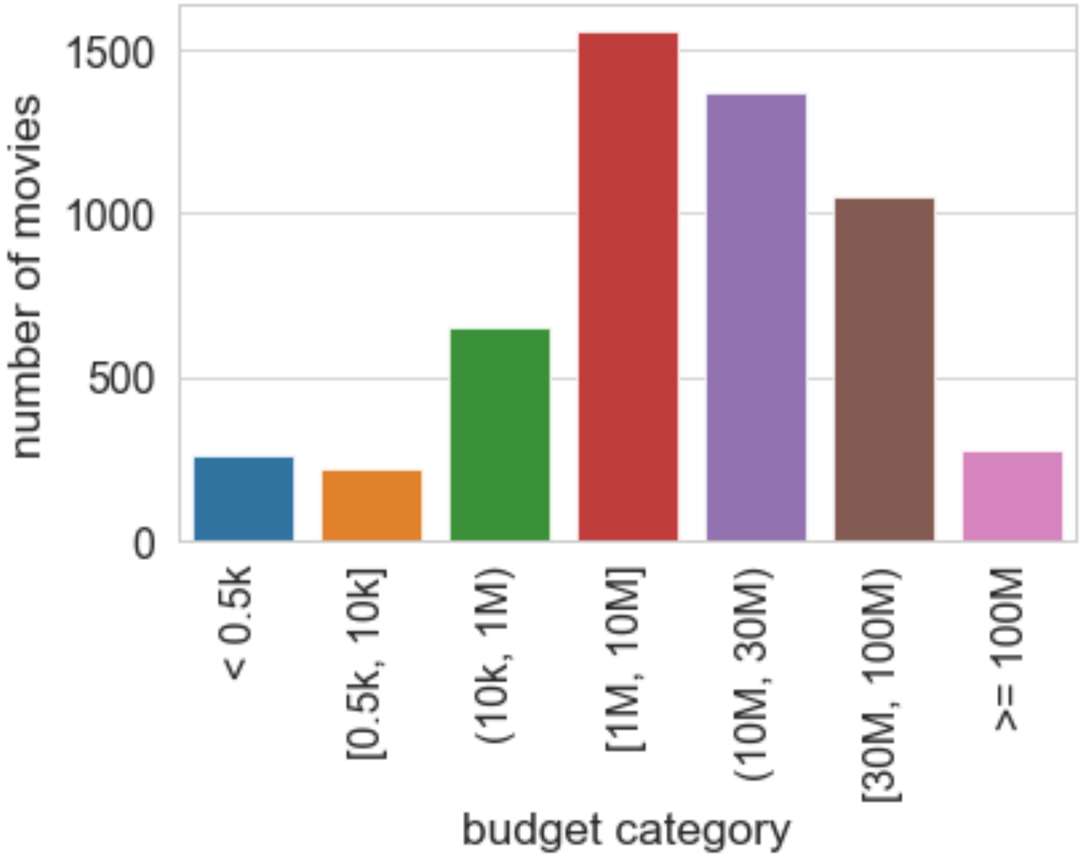
	title	first_genres	first_production_countries	first_production_companies
count	5247	5247	5247	5247
unique	5177	20	82	2255
top	The Three Musketeers	Drama	United States of America	missing
freq	3	1266	2891	431

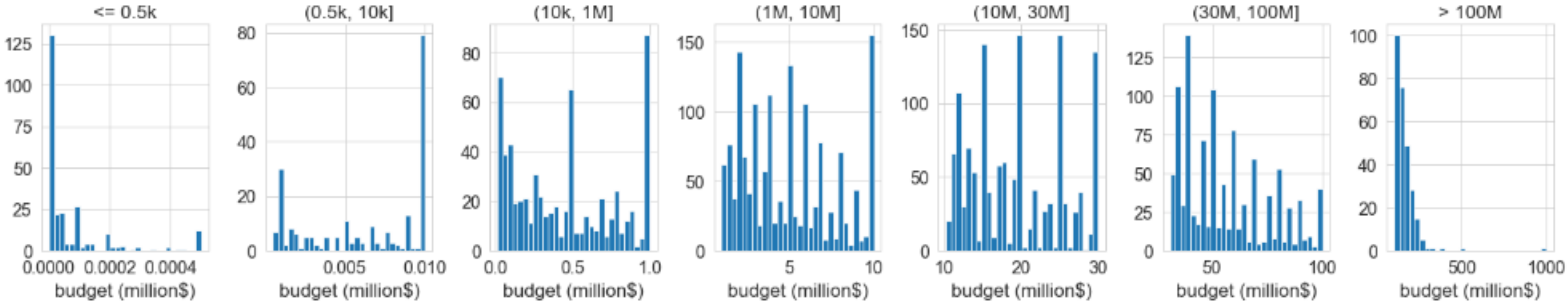
```
Remove 2.65% of the data.
```

Extra: Categorize the budget variable



	< 0.5k	[0.5k, 10k]	(10k, 1M)	[1M, 10M]	(10M, 30M)	[30M, 100M)	>= 100M
budget_category	224	206	586	1529	1373	1050	278


number of movies
< 0.5k
[0.5k, 10k]
(10k, 1M)
[1M, 10M]
(10M, 30M)
[30M, 100M)
>= 100M
budget category


<= 0.5k
budget (million$)
(0.5k, 10k]
budget (million$)
(10k, 1M]
budget (million$)
(1M, 10M]
budget (million$)
(10M, 30M]
budget (million$)
(30M, 100M]
budget (million$)
> 100M
budget (million$)