Course: Data Mining

Homework 1

Student ID: 20C14001

Student Name: Le Duong Tuan Anh

Dataset: "top_movies_by_title.csv"

1) Desbribe the dataset

Dataset type: Record.

This dataset has 200 rows and 5 columns.

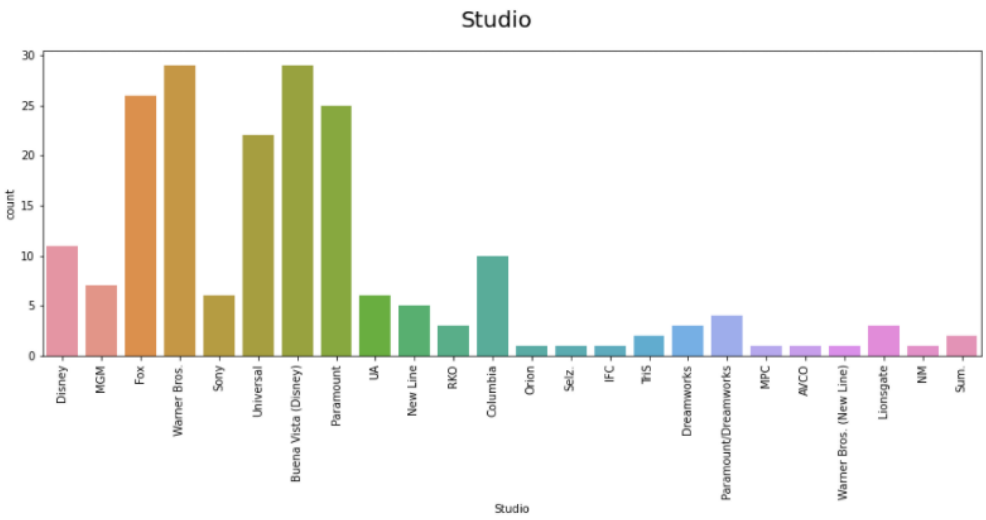
Column	Type	Datatype	Has missing value?
Title	Metadata	String	No
Studio	Nominal	String	No
Gross	Interval-scaled	Integer	No
Gross (Adjusted)	Interval-scaled	Integer	No
Year	Quantitative (continuous interval)	Integer	No

2) Appy basic statictical descriptions for the dataset

2.1. Studio

	Studio	Count
0	AVCO	1
1	Buena Vista (Disney)	29
2	Columbia	10
3	Disney	11
4	Dreamworks	3
5	Fox	26
6	IFC	1
7	Lionsgate	3
8	MGM	7
9	MPC	1
10	NM	1
11	New Line	5
12	Orion	1
13	Paramount	25
14	Paramount/Dreamworks	4
15	RKO	3
16	Selz.	1
17	Sony	6
18	Sum.	2
19	TriS	2
20	UA	6
21	Universal	22
22	Warner Bros.	29
23	Warner Bros. (New Line)	1

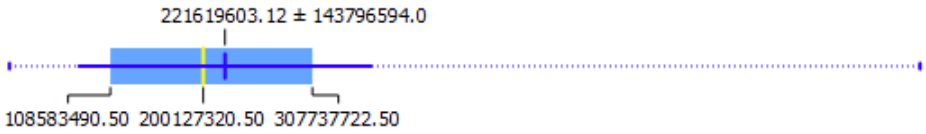
Distribution Chart

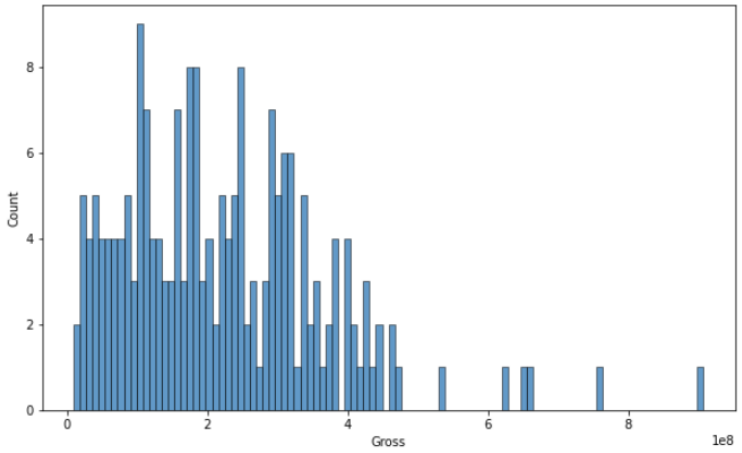


Tag Cloud

2.2. Gross

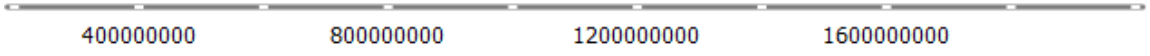
	count	mean	std	min	25%	50%	75%	max
Gross	200.00	221619603.12	144157439.17	9183673.00	108782382.75	200127320.50	306953495.25	906723418.00

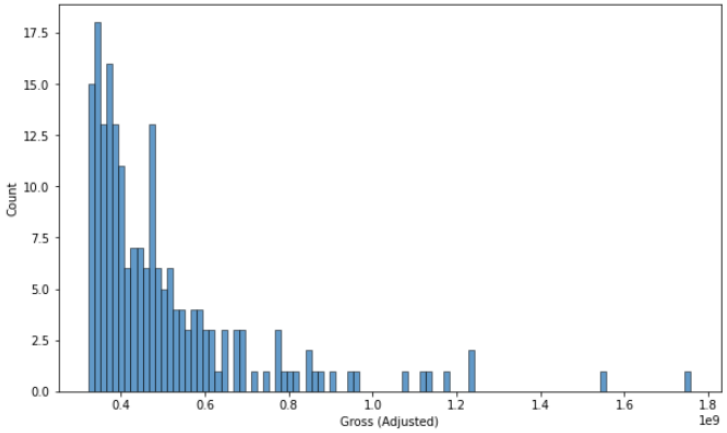




2.3. Gross (Adjusted)

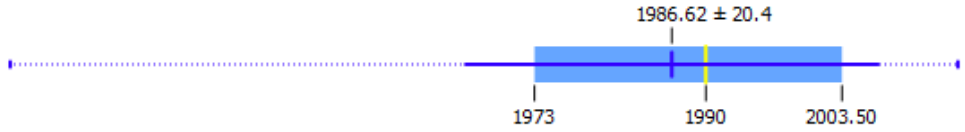
	count	mean	std	min	25%	50%	75%	max
Gross (Adjusted)	200.00	504198276.50	215981430.31	322261900.00	367780375.00	438857000.00	551213100.00	1757788200.00

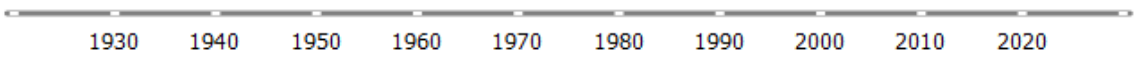


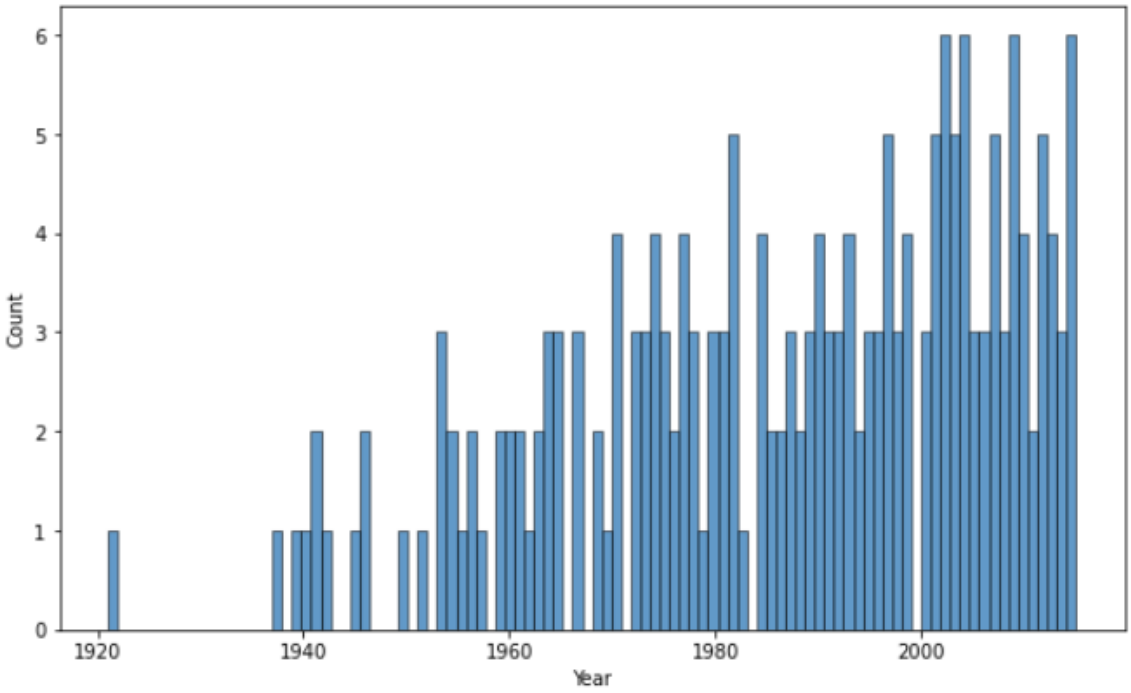


2.4. Year

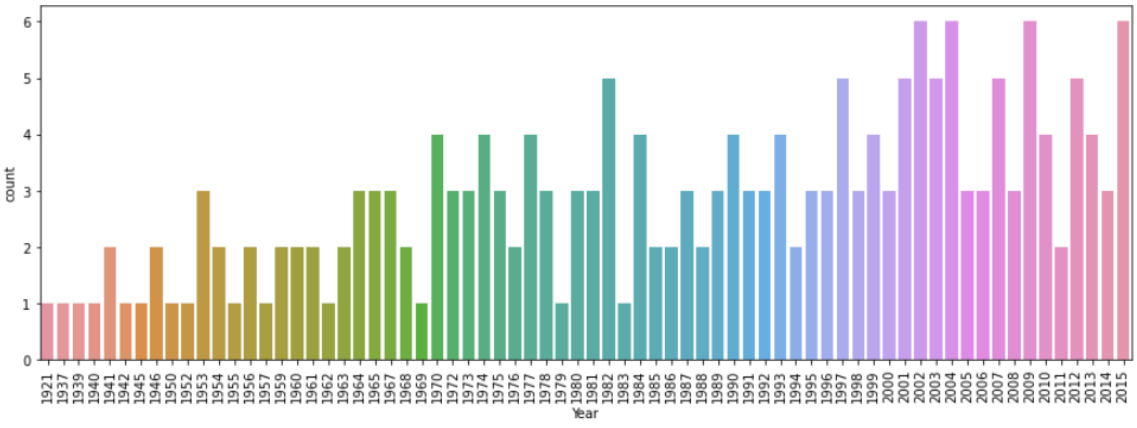
	count	mean	std	min	25%	50%	75%	max
Year	200.00	1986.62	20.49	1921.00	1973.00	1990.00	2003.25	2015.00





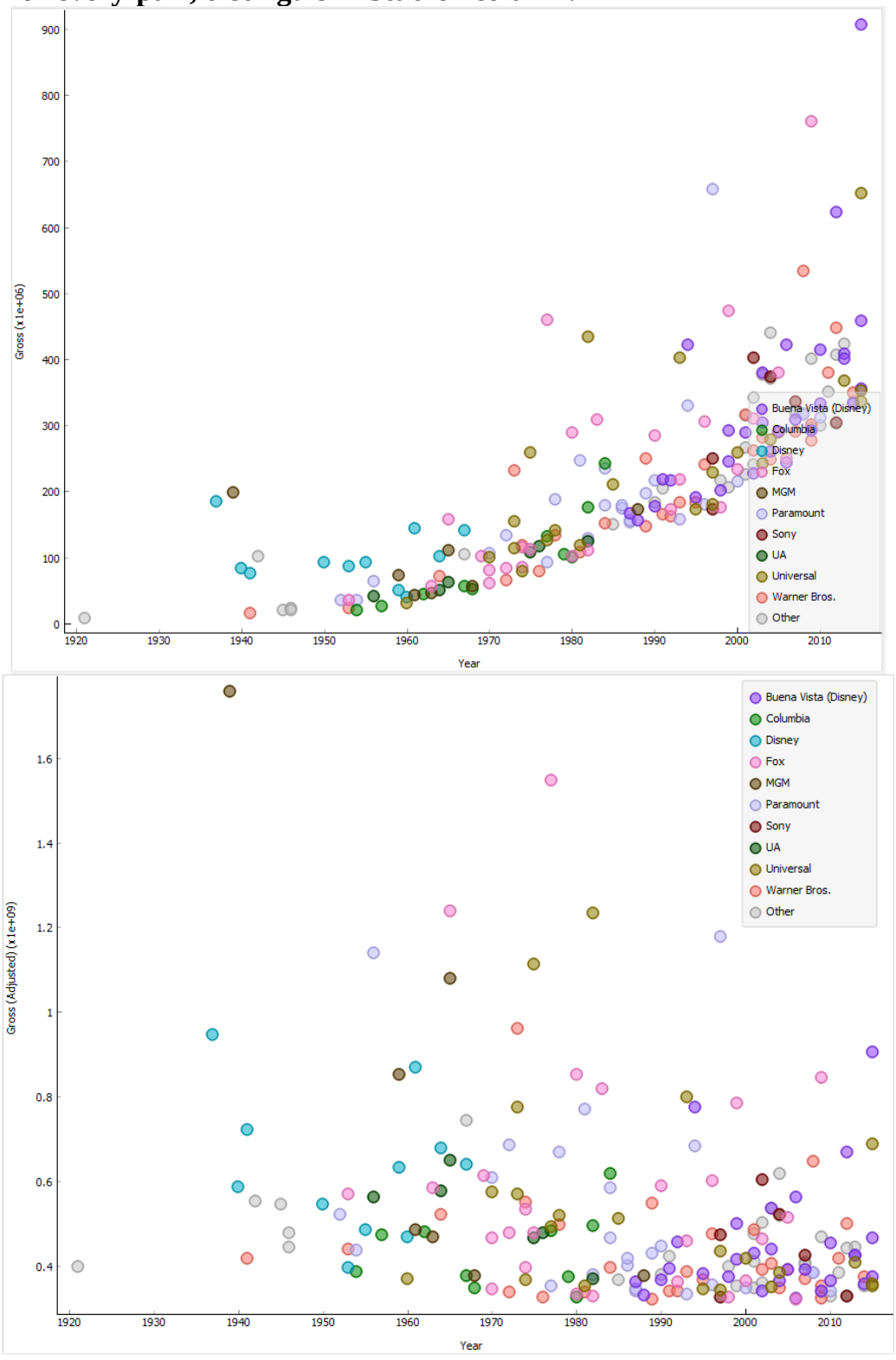


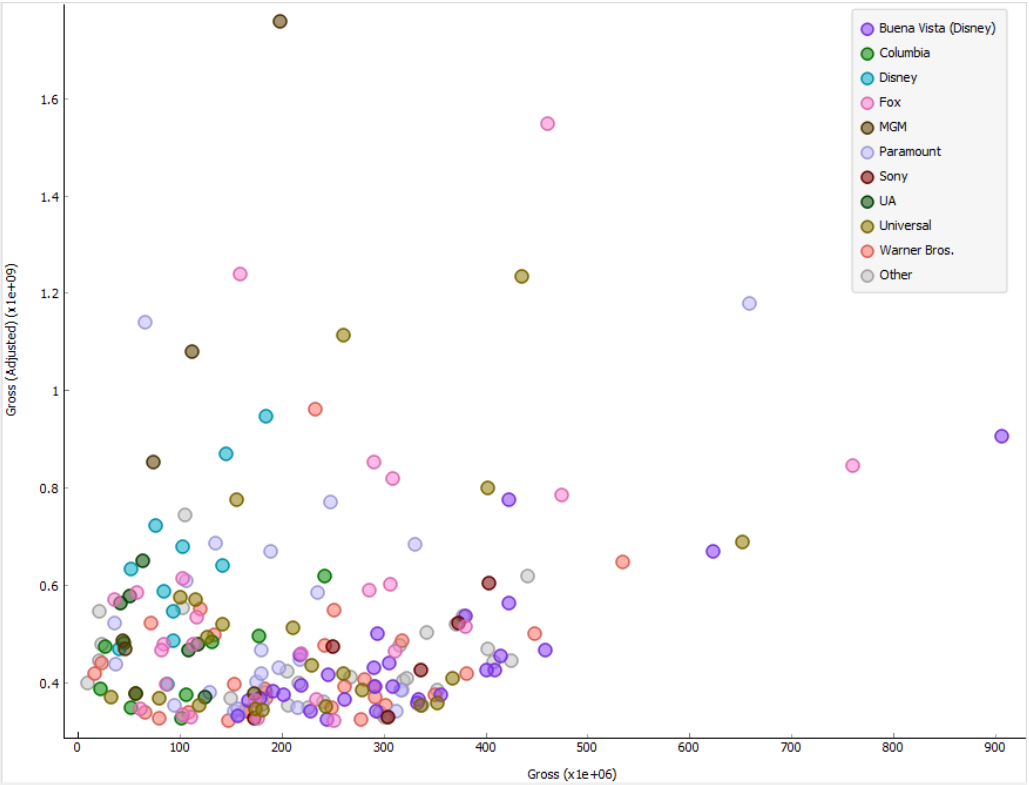
Year



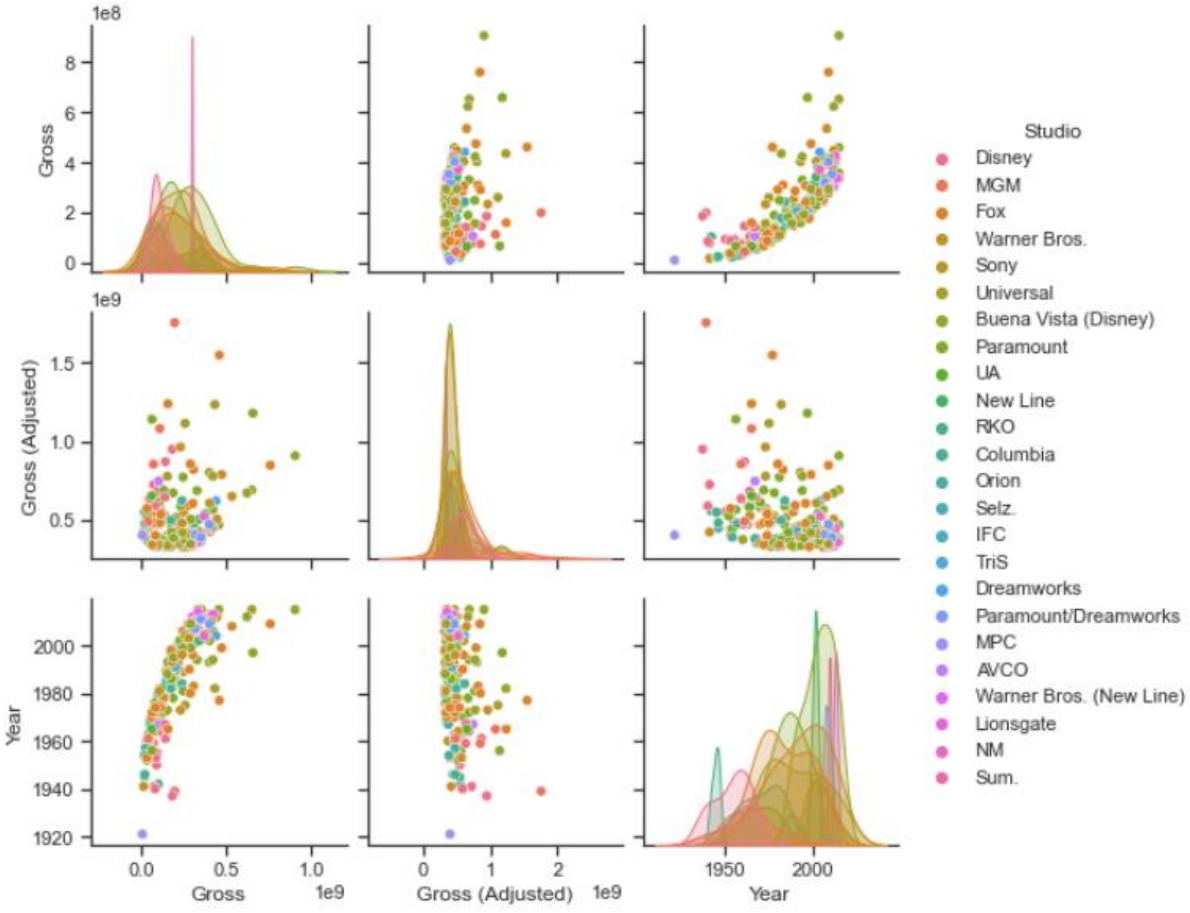
3) Visualize this dataset by using scatterplot matrix

3.1. For every-pair, distinguish "Studio" column.





3.2. For whole dataset, distinguish "Studio" column.



3.2. For whole dataset, without distinguish “Studio”.