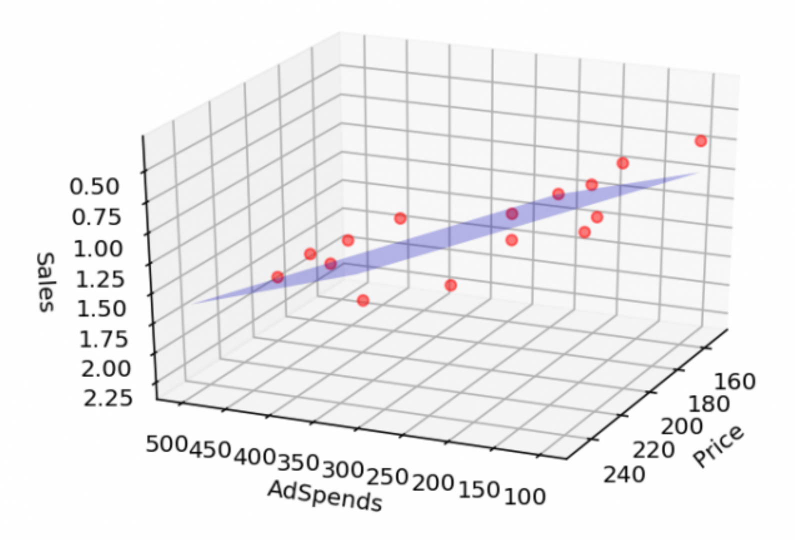
Visualizing Data for Movie Audience Expectation

Team Members: Jongwon Lee (1)

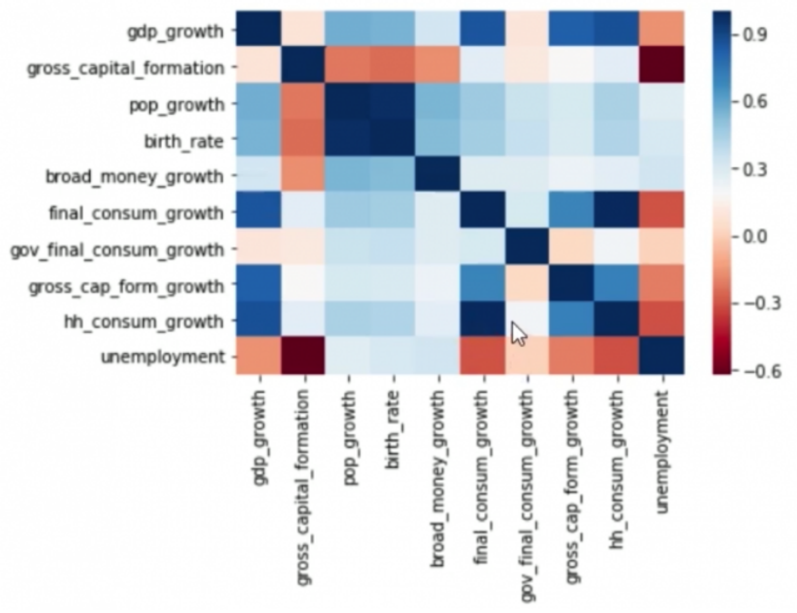
Abstract: This project will visualize data collected for indie-film audience expectation done with multiple linear regression. It is a undergraduate level paper that already had been written and got awarded at Sungkyunkwan University in Korea. (I was a member of the writers) The abstract of the thesis will be provided in English. It collected data and did a multiple regression to predict audience turnout for indie-film that were imported in Korea.

Question: Can we visualize the result of multiple linear regression with more than 3 independant variables? If impossible, we can compare each variable with 3 xyz. For example, like the following with each variable.



Objectives: This project will provide a heatmap with seaborn to visualize correlation of data, when we decide to remove variables.

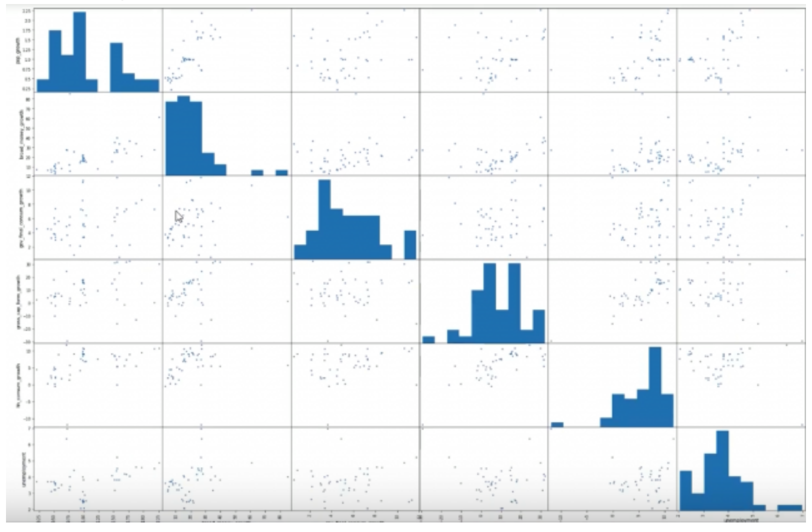
For example, like the following with all variables.



dark blue(1) square with two different variable states two are related, so one should be dropped.

Motivation is that with just numerical results provided by libraries or stata, unless the performer is extremely familiar to the regression process or principle, it seems hard to feel what each values mean. For example, when I write a paper with a bunch of statistical results but no visualization, it is hard for the general readers to grasp the idea.

Also, it can be visualized like the following with multiple scatterplot.



we want each square to be scattered not having a pattern.

These two charts are from this source: https://www.youtube.com/watch?v=8DhvVs59It4

On top of that, I would like to create interactive graph for predicting the result of a new input.

Datasets and Methods: I have collected Dataset in 2016 having 300 movies with 6 independent variables. At that time, I have dropped a few variables because of their p-value and correlation. I will be adding more variable with existing movies and see if I can improve the regression.

Multiple Linear Regression will be done with scikit-learn. Visualization with seaborn, pandas and etc.

References: https://www.youtube.com/watch?v=8DhvVs59It4 ,
https://drive.google.com/drive/folders/0B0hBEl2qMK06OGpNZktRMk9tS2s

In [3]:
```
import pandas as pd
df = pd.read_csv("movie_list.csv")
df
```

Out[3]:

	id	code	title_en	title_ko	released_on	country	fscreen	preview	ani	naver	youtube	fweek	expectati
0	13	20149629	Begin Again	비긴 어게인	2014-08-13	미국	185	9215	0	1276	94057	133628	131589.7320
1	154	20150020	Son of Saul	사울의 아들	2016-02-25	헝가리	47	3007	0	381	6111	11292	11492.8755
2	107	20147684	Night Train to Lisbon	리스본행 야간열차	2014-06-05	독일	52	4584	0	431	3097	19679	19301.9311
3	33	20166721	Foosball	장난감이 살아있다	2016-09-07	아르헨티나	212	1390	1	976	7075	67462	66008.0482
4	247	20149860	Brave Rabbit	브레이브 래빗 : 새로운 영웅의 탄생	2014-08-28	중국	38	415	1	40	70	4829	4935.1340
...	...	...	...	...	...	...	...	...	...	...	...	...	
294	220	20159965	Yowamushi Pedal the Movie	겁쟁이 페달: 더 무비	2016-01-14	일본	52	679	1	1210	20334	6197	31829.6207
295	291	20147723	Twice Born	투와이스 본	2014-10-30	이탈리아	30	328	0	144	1349	3380	-10622.0739
296	294	20154484	Our Last Tango	라스트 탱고	2015-12-31	독일	19	775	0	153	239	3222	-11700.9490
297	270	20144881	Journey to the West: Conquering the Demons	서유기 : 모험의 시작	2015-02-05	중국	68	1575	0	1436	12113	4103	23200.0837
298	279	20140426	Vijay and I	나의 첫 번째 장	2014-09-11	벨기에	17	126	0	173	365	3754	-15098.1529