다변량분석 과제1_2020170816 김소민

[Q1] 본인이 스스로 Multivariate Linear Regression 모델을 적용하는데 적합한 데이터셋을 선정하고 선정 이유를 설명하시오. 데이터셋 탐색은 아래에서 제시된 Data Repository를 포함하여 여러 Repository를 검색해서 결정하시오. 보고서에는 데이터를 다운로드할 수 있는 링크를 반드시 제공하시오.

⇨ 해당 과제를 수행할 때 가장 중요하게 생각한 점이 '내가 관심이 있는 분야에서 흥미로운 결과를 얻을 수 있는가?'였다. 그래서 가장 관심이 있는 분야이며 향후 진로와도 관련이 있는 게임 쪽의 데이터를 찾기로 했다.

처음에 찾은 데이터셋은 오버워치 프로게이머 선수들의 마우스 감도 데이터였다. 선수들의 닉네임과 역할군, 메인 영웅, dpi, 감도, edpi 등의 자료를 통해 선수의 역할군이 선호 감도에 어느 정도 영향을 주는 지를 측정하는 것이 목표였다. 하지만 얼마 지나지 않아 해당 데이터셋이 Multivariate Linear Regression을 적용하는 데 적합하지 않다는 것을 깨달았다. 게임 내부에서의 감도를 결정하는 것은 edpi인데, 이는 단순히 dpi와 감도의 곱셈으로 나오는 값이었다는 것이다. 즉, 단순히 Y= X1 * X2의 답이 정해져 있는 문제가 될 수밖에 없었던 것이다. 이를 해결하기 위해 dpi와 감도 중 하나의 값을 빼려고 했지만, cmPer360 값 역시 edpi에 정확히 반비례하는 값이었다. 이런 이유들을 종합하여 Multicollinearity가 너무 커지고, 설명변수의 수가 지나치게 적었다는 점을 감안하여 해당 데이터셋을 폐기하기로 하였다.

이후에는 종속변수가 설명변수의 단순 연산만으로 구해지지 않고, 다양한 설명변수를 가지는 데이터셋을 찾는 데 집중하였다. 결국 게임 분야에서 찾을 수 있던 데이터는 Video Game Sales with Ratings, 즉 평점에 따른 게임 판매량이었다. 종속변수를 판매량으로 정하고, User Score와 Critic Score 등을 설명변수로 정하면 원하는 회귀식을 얻을 수 있을 것이라 판단했다.

https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings?resource=download

[Q2] 해당 데이터셋의 종속변수와 설명변수는 어떤 것들이 있는가? 분석 전에 아래 세 가지 질문에 대해서 스스로 생각해보고 답변을 하시오.

⇨ 우선, 인덱스에 해당하는 Name은 분석과 무관하므로 제외한다. 또한, 해당 데이터셋에서 국가별 판매량을 알려는 것이 아니므로 Global_Sales를 종속변수로 정하고 나머지 Sales는 제외한다(NA_Sales, EU_Sales, JP_Sales, Other_Sales). 한편 데이터를 살펴본 결과 Developer 칼럼에 복수의 회사명이 들어갈 수 있다는 것을 알게 되었다. 이후 데이터를 전처리 할 때 명목형 데이터에 대해 1-of-C coding을 수행할 예정인데, 이 때문에 하나의 칼럼에 복수의 회사 이름이 들어가면 제대로 된 분석을 할 수 없을 것이라고 판단했다. 다음 도표가 이를 명확하게 보여준다.

	Name	Developer
36	Call of Duty: Modern Warfare 2	Infinity Ward
37	Call of Duty: Modern Warfare 3	Infinity Ward, Sledgehammer Games

1-of_C 코딩을 하면 "Infinity Ward"와 "Infinity Ward, Sledgehammer Games"가 별개의 칼럼으로 저장이 된다. 즉, Call of Duty: Modern Warfare 3의 Developer가 "Infinity Ward, Sledgehammer Games"라는 새로운 하나의 회사로 간주되는 것이다. 이렇게 된다면 모든 공동 제작된 게임의 Developer 정보가 사실상 유실되는 것이므로, 이를 방지하기 위해 Developer를 설명변수에서 제외하기로 했다. 따라서 최종적인 변수 목록은 다음과 같다. 쉽게 알아볼 수 있도록 종속변수에 형광색 표시를 해두었다.

[Platform, Year_of_Release, Genre, Publisher, Global_Sales, Critic_Score, Critic_Count, User_Score, User_Count, Rating]

1. 이 데이터는 종속변수와 설명변수들 사이에 실제로 "선형 관계"가 있다고 가정할 수 있겠는가? 가정할 수 있음/없음 판단에 대한 본인의 생각을 서술하시오.

⇨ 스스로가 게임을 좋아하는 사람의 입장에서, 어떤 게임을 살 때에는 다양한 요소들을 고려한다. 필자가 중점적으로 보는 요소에는 사람들의 평가, 게임 플랫폼의 종류, 게임의 장르 등이 있다. 다른 게이머들이 게임을 구매할 때 고려하는 요소도 여기서 크게 벗어나지 않을 것이라는 것을 경험을 통해 추론하였다. 이러한 점과 해당 데이터셋에 언급한 요소들이 다수 포함되어 있다는 점을 근거로 하여, 일부 특수한 케이스를 제외하면 종속변수와 설명변수들 사이에 전반적으로 "선형 관계"가 존재한다고 판단하였다.

2. 이 데이터에서 제공된 설명변수들 중에서 높은 상관관계가 있을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?

⇨ Critic_Score와 User_Score는 게임에 대한 비평가들과 일반 유저들의 평점이 담겨있으므로 판매량에 직결될 것이다. 만족도의 지표인 평점이 높으면 더 많은 사람들이 평점을 보고 게임을 구매하게 될 것이기 때문이다. Critic_Count와 User_Count는 많을수록 게임의 유명세를 나타내므로 종속변수에 상당히 큰 영향을 끼칠 것으로 예상된다. 게임의 장르(Genre)는 해당 분야의 대중성과 큰 관련이 있으므로 판매량과 상당히 높은 상관관계가 있을 것이라고 추정된다.

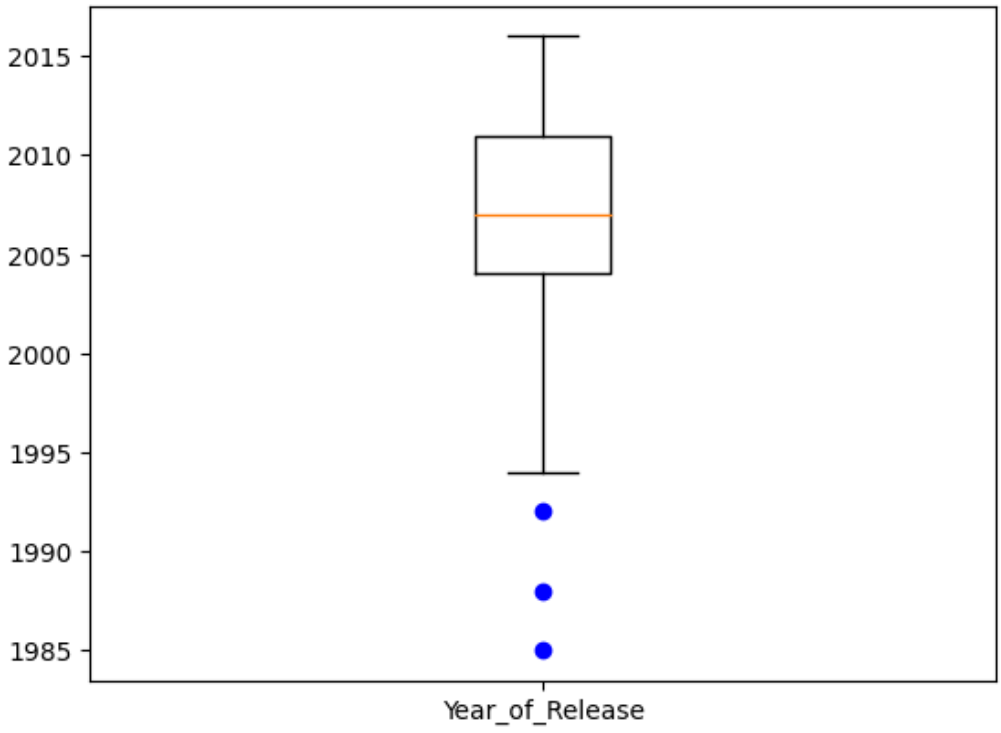
3. 제공된 설명변수들 중에서 종속변수를 예측하는데 필요하지 않을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?

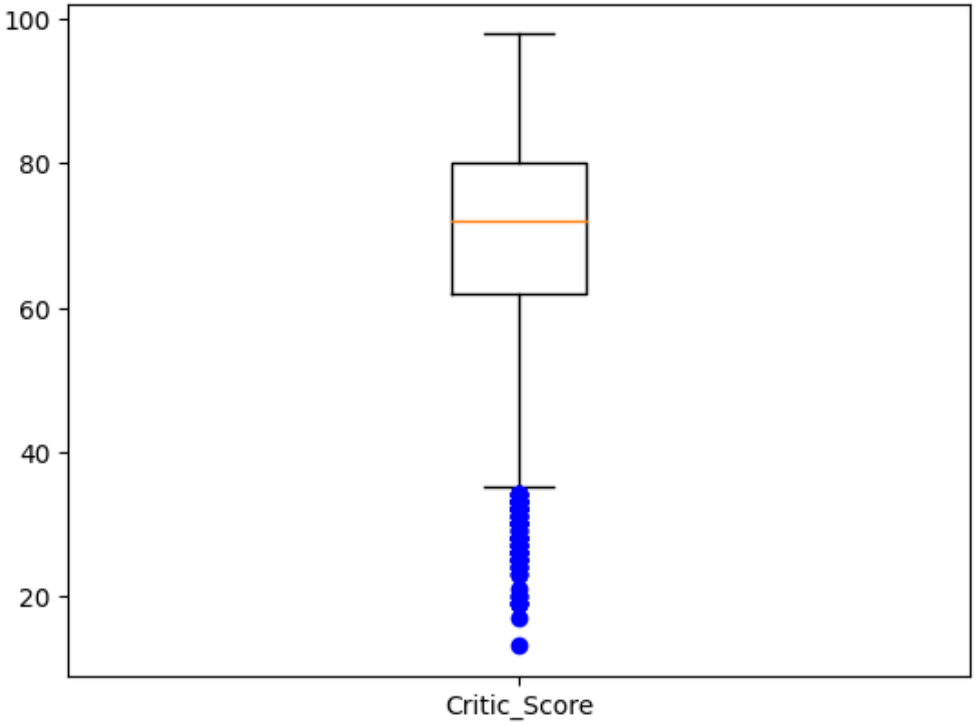
⇨ 발매 년도(Year_of_Release)는 종속변수를 예측하는데 필요하지 않을 것이다. 어느 정도 게임을 즐기는 사람들은 출시일과 무관하게 재밌어 보이는 게임을 구매하기 때문이다. 게임의 등급(Rating)은 해당 게임이 어느 나이대의 사람에게 적합한 지를 나타내는 척도이다. 게임을 즐기는 사람들의 나이대가 다양하고, 이들이 즐기는 게임의 성향은 개인차가 클 것이기 때문에 등급이 게임 판매량에 미치는 영향은 너무나 복합적으로 작용될 것이라고 보인다. 지나치게 복잡한 관계성은 오히려 무관한 것과 구분하기 힘들 것이라 생각하여 Rating 또한 판매량을 예측하는 데 필요 없을 것이라고 판단했다.

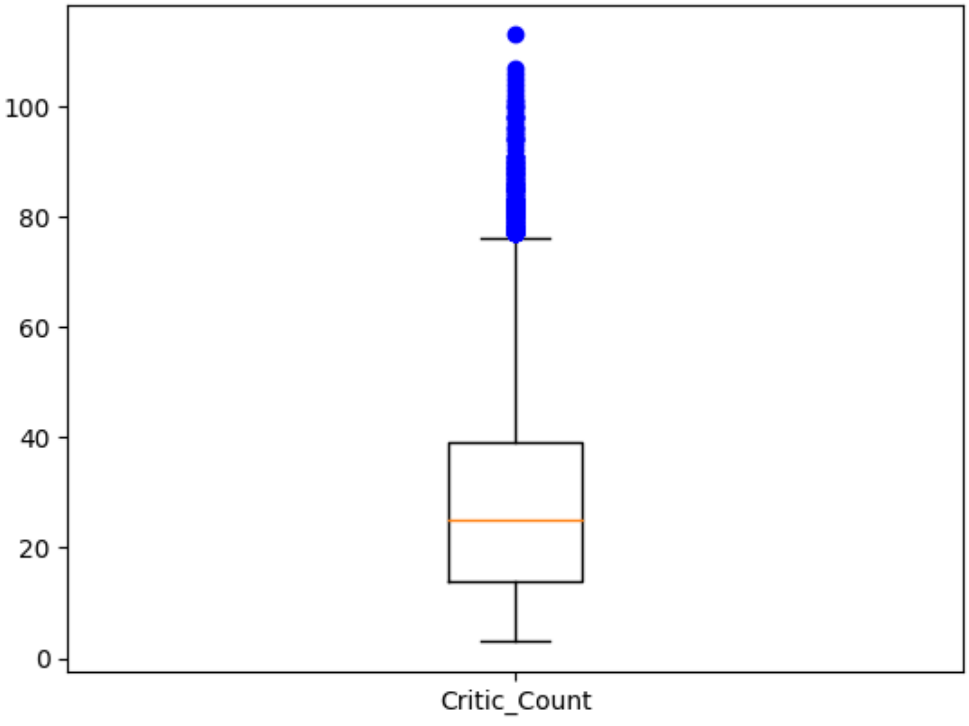
[Q3] 개별 입력 변수들에 대하여 각각 다음과 같은 단변량 통계량을 계산하고 Box plot을 도시하시오: Mean, Standard deviation, Skewness, Kurtosis. 전체 변수 중에서 정규분포를 따른다고 할 수 있는 변수들은 몇 개인가? 정규분포를 따른다고 가정한 근거는 무엇인가?

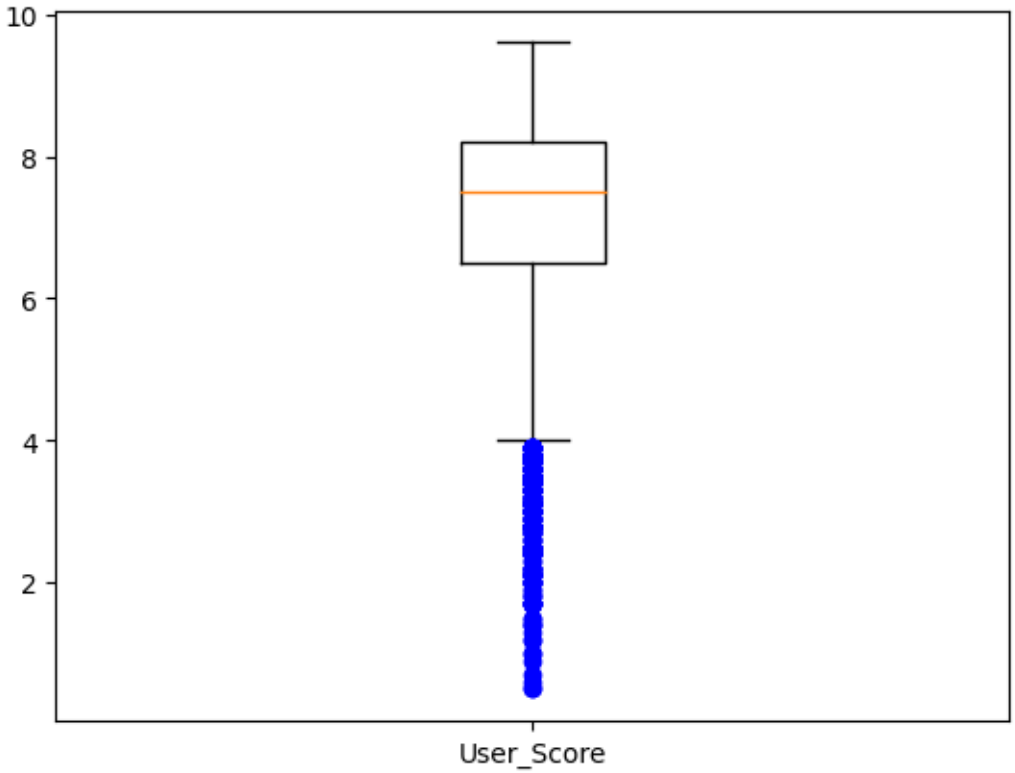
⇨ 단변량 통계량은 명목형 변수를 제외한 변수들에서만 계산이 가능하므로, 해당 과정은 다음의 변수들에서만 수행되었다: [Year_of_Release, Critic_Score, Critic_Count, User_Score, User_Count]

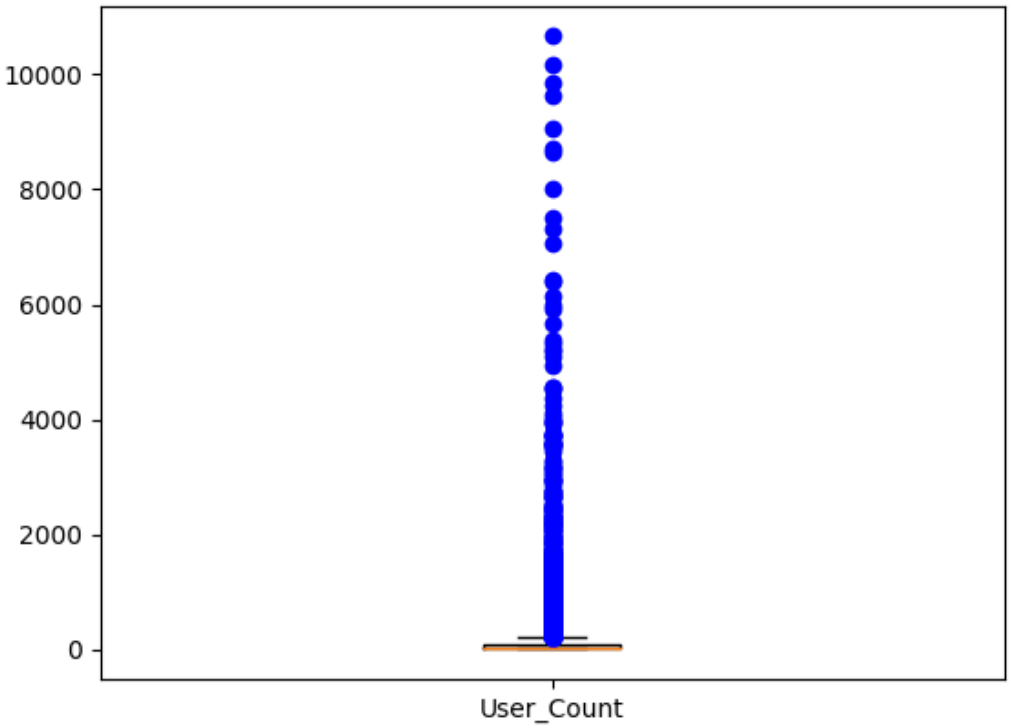
	Attribute	mean	std	skewness	kurtosis
0	Year_of_Release	2007.436777	4.211248	-0.518370	0.110819
1	Critic_Score	70.272088	13.868572	0.394749	-0.745604
2	Critic_Count	28.931136	19.224165	0.732260	1.030652
3	User_Score	7.185626	1.439942	1.610710	-1.219081
4	User_Count	174.722344	587.428538	103.096429	8.665041




100
80
60
40
20
Critic_Score


100
80
60
40
20
0
Critic_Count





Kline(2005)에 따르면 Skewness의 절댓값이 3보다 작고, Kurtosis의 절댓값이 8보다 작으면 정규 분포를 따른다. 해당 데이터셋에서 이러한 기준을 적용한다면 User_Count를 제외한 4개의 변수들이 정규분포를 따른다고 할 수 있다.

[Q4] [Q3]의 Box plot을 근거로 각 변수들에 대한 이상치(너무 값이 크거나 작은 객체) 조건을 정의하고, 해당하는 객체들을 데이터셋에서 제거해 보시오.

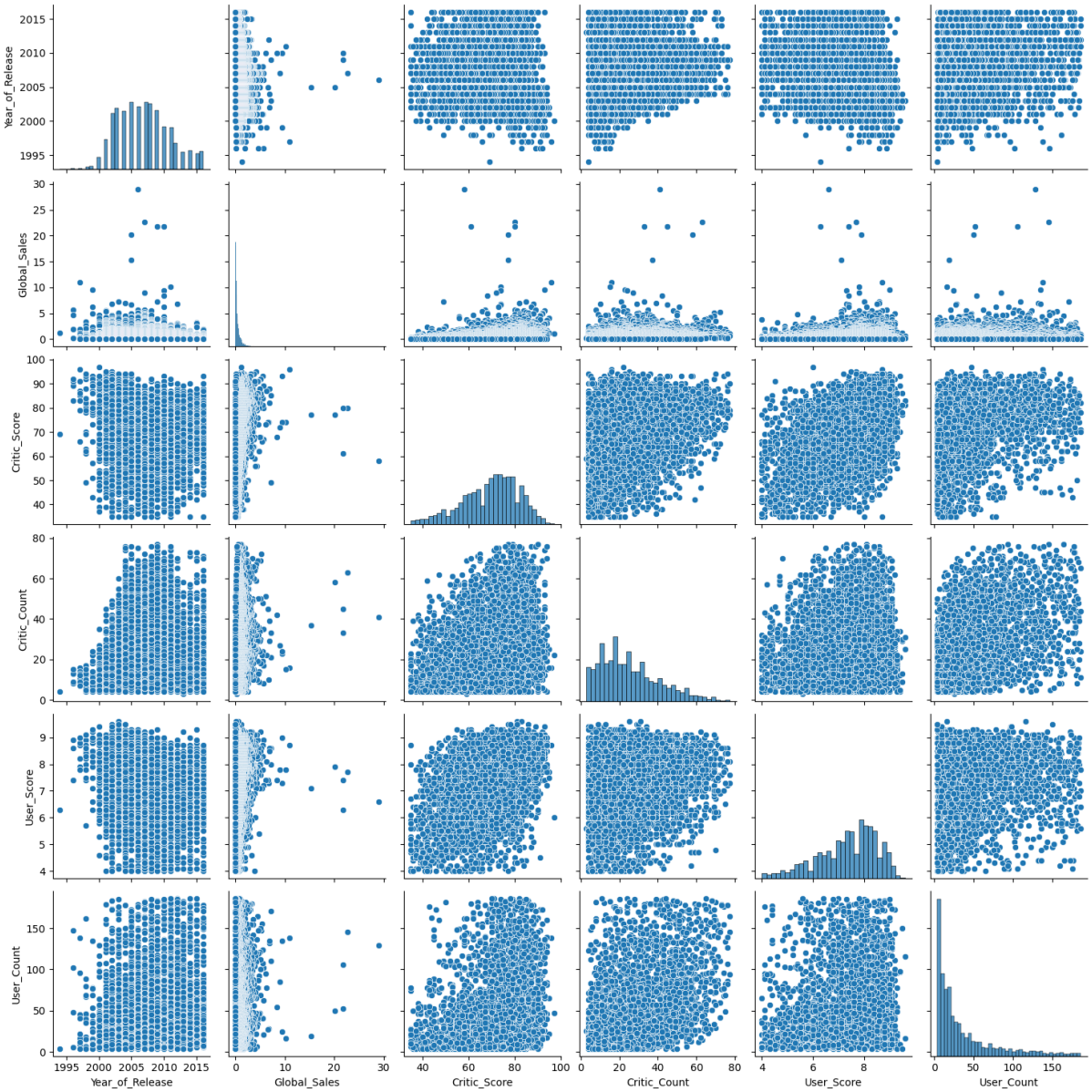
⇨ Box plot에서 whiskers 밖에 존재하는 값들을 이상치라고 부른다. Whisker의 범위는 [Lower quartile(q1) - 1.5 IQR, Upper quartile(q3) + 1.5 IQR]이기 때문에, 이상치 조건은 q1 – 1.5 IQR보다 작거나 q3 + 1.5 IQR보다 큰 값이 된다. 해당하는 객체들을 데이터셋에서 제거해보면 tuple수가 6825에서 5504으로 줄어든 것을 알 수 있다.

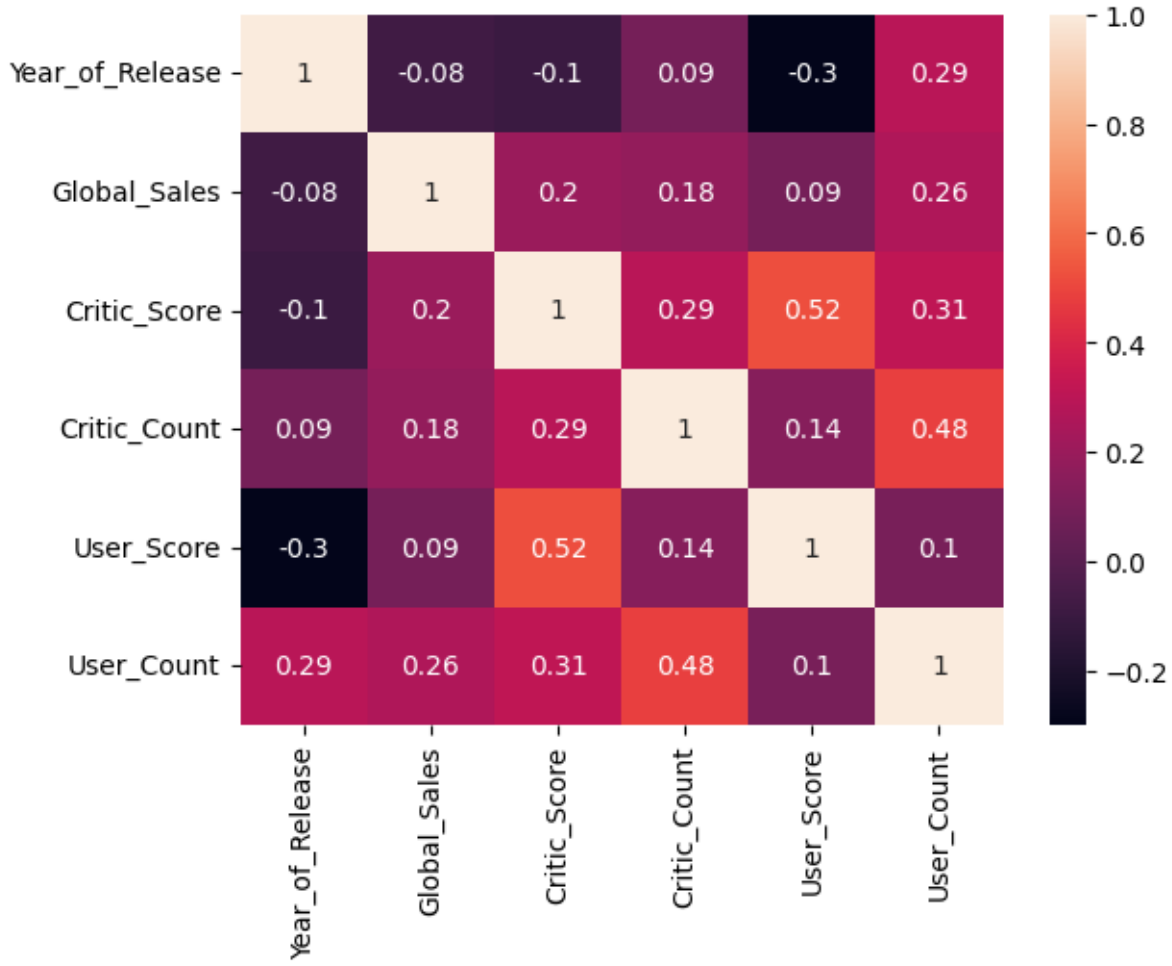
	Platform	Year_of_Release	Genre	Publisher	Global_Sales	Critic_Score	Critic_Count	User_Score	User_Count	Rating
7	Wii	2006.0	Misc	Nintendo	28.92	58.0	41.0	6.6	129.0	E
13	Wii	2007.0	Sports	Nintendo	22.70	80.0	63.0	7.7	146.0	E
14	X360	2010.0	Misc	Microsoft Game Studios	21.81	61.0	45.0	6.3	106.0	E
15	Wii	2009.0	Sports	Nintendo	21.79	80.0	33.0	7.4	52.0	E
19	DS	2005.0	Misc	Nintendo	20.15	77.0	58.0	7.9	50.0	E
...	...	...	...	...	...	...	...	...	...	...
16634	XOne	2016.0	Racing	Milestone S.r.l	0.01	63.0	8.0	8.2	22.0	E
16656	WiiU	2016.0	Action	Nintendo	0.01	81.0	46.0	8.5	151.0	E
16677	GBA	2002.0	Fighting	Midway Games	0.01	81.0	12.0	8.8	9.0	M
16700	PC	2011.0	Shooter	Destineer	0.01	61.0	12.0	5.8	43.0	T
16706	PC	2011.0	Strategy	Unknown	0.01	60.0	12.0	7.2	13.0	E10+

5504 rows × 10 columns

다음 각 물음에 대해서는 [Q4]에서 제거된 객체들을 제외시킨 데이터프레임을 사용하여 답하시오.

[Q5] 가능한 모든 두 쌍의 입력 변수 조합에 대한 산점도(scatter plot)등을 도시하여 입력변수 간 상관성에 대한 분석을 수행해 보시오. 어떤 두 조합의 변수들이 서로 강한 상관관계가 있다고 할 수 있는가? 이렇게 강한 상관관계가 발생한 변수들은 상식적으로도 상관관계가 높은 변수들이라고 할 수 있는가?





⇨ Scatter plot의 모양과 correlation의 절댓값에서 Critic_Score와 User_Score, 그리고 Critic_Count와 User_Count의 변수들이 서로 상당히 강한 상관관계가 있다는 결론을 내릴 수 있다.

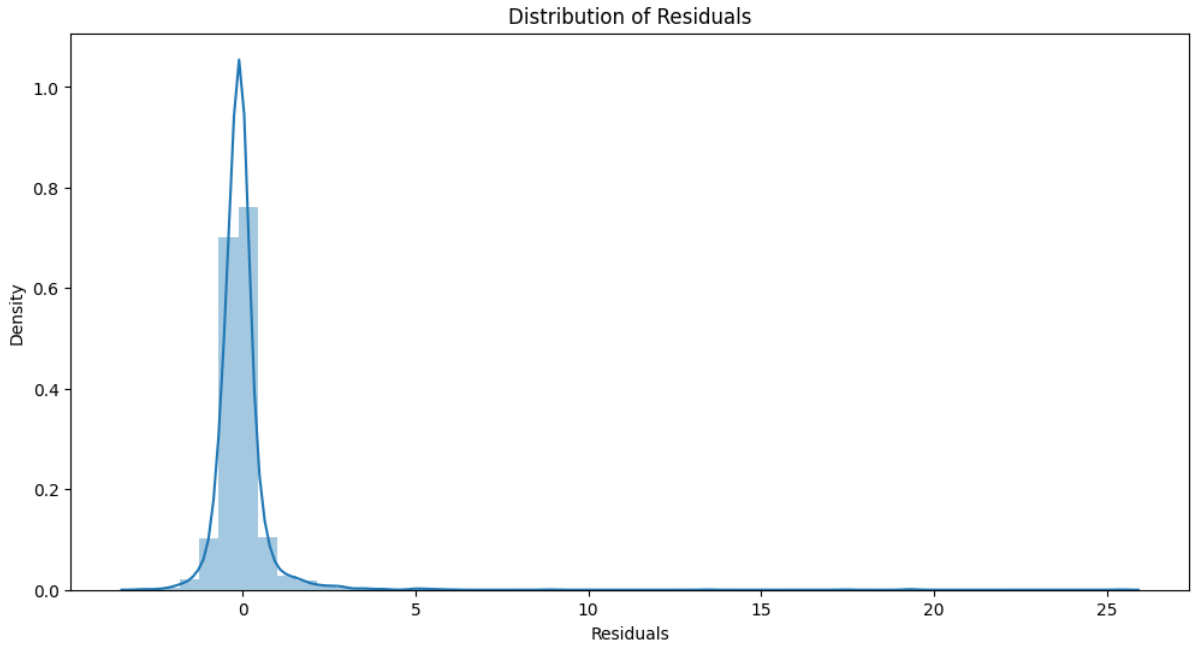
이러한 상관관계가 상식과도 부합하는 지를 판단하기 위해 우선 Critic_Score와 User_Score를 살펴보겠다. 게임 자체의 완성도가 높으면 비평가와 일반 유저 모두 높은 점수를 주고, 게임 자체가 졸작이라면 비평가와 일반 유저가 모두 낮은 점수를 줄 것이므로 이 두 변수는 상관관계가 높은 변수라고 할 수 있을 것이다. 한편, Critic_Count와 User_Count는 게임의 유명세와 큰 관련이 있을 것이다. 유명한 게임의 경우 평점을 남긴 비평가의 수와 일반 유저의 수가 모두 많을 것이고, 유명하지 않은 게임이라면 적은 수의 비평가와 일반 유저들이 평점을 남길 것이기 때문이다. 이를 토대로 해당 변수 조합들의 상관관계가 상식적으로도 납득이 가능하다는 판단을 내릴 수 있다.

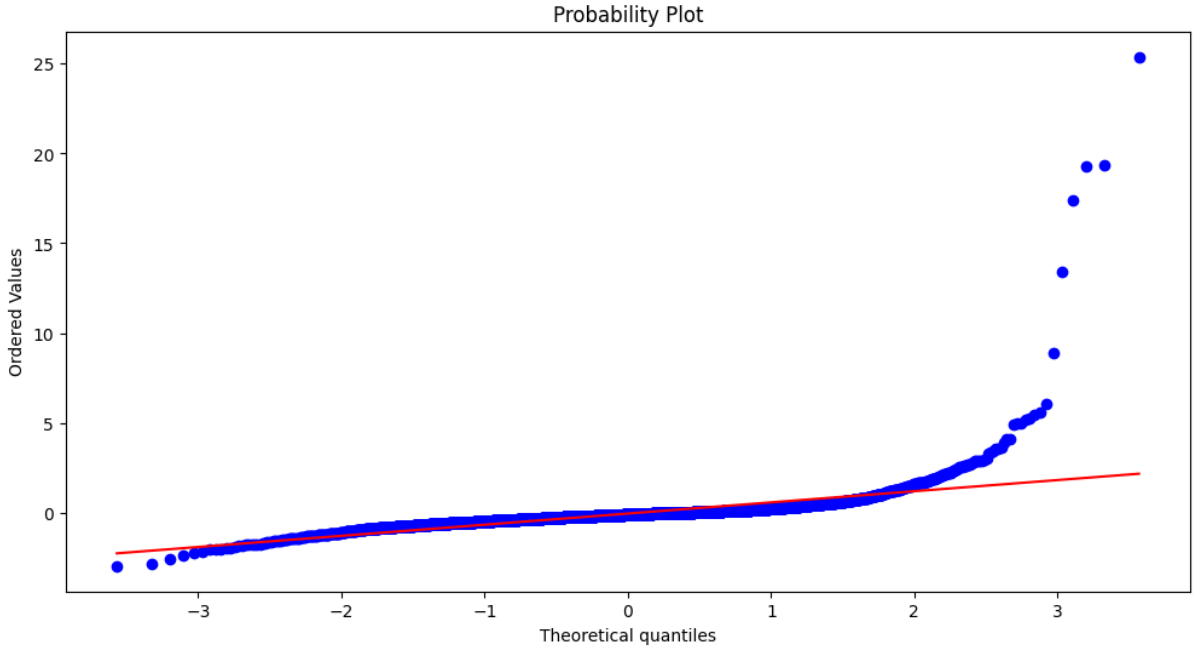
[Q6] 전체 데이터셋을 70%의 학습 데이터와 30%의 테스트 데이터로 분할한 후 모든 변수를 사용하여 MLR 모델을 학습한 뒤, Adjusted R2값을 통해 데이터의 선형성(linearity)을 판단해 보시오. Residual plot 과 Q-Q Plot을 도시하고 Ordinary Least Square 방식의 Solution이 만족해야 하는 가정들이 만족될만한 수준인지 정성적으로 판단해 보시오.

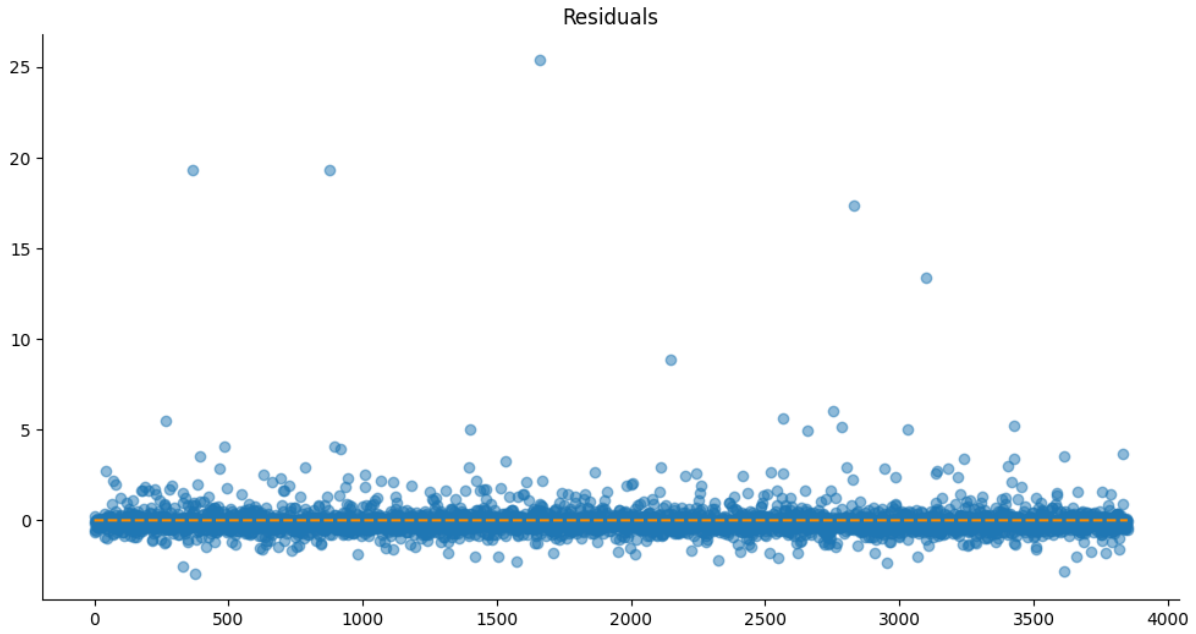
```
# Calculate Adjusted R_squared value
adr2 = 1-(1-mlr.score(x_trn,y_trn))*(x_trn.shape[0] - 1)/(x_trn.shape[0] - x_trn.shape[1]- 1)
adr2
```

```
0.18972278998762504
```

⇨ Adjusted R2값은 0.190로, 매우 낮은 편에 속해 데이터가 선형성을 지니지 않는다는 것을 나타낸다.







⇨ Residual plot과 Q-Q plot의 모양을 통해 해당 선형 회귀 모델이 정규성 가정을 만족한다는 것을 알 수 있다.

```
Assumption : Little to no multicollinearity among predictors
Year_of_Release: 8394.008525371462
Critic_Score: 69.58666491817594
Critic_Count: 7.991191978317888
User_Score: 70.50738660243482
User_Count: 3.6883747105731395
Wii: 6.282285679667283
```

.

.

.

```
54 cases of possible multicollinearity
17 cases of definite multicollinearity

Assumption not satisfied

Coefficient interpretability will be problematic
Consider removing variables with a high Variance Inflation Factor (VIF)
```

⇨ 상당히 많은 변수가 multicollinearity를 보여 해당 assumption test를 만족하지 못한 것을 알 수 있다.

[Q7] 유의수준 0.01에서 모형 구축에 통계적으로 유의미한 변수들은 어떤 것들이 있는가? 해당 변수들은 종속변수와 양/음 중에서 어떤 상관관계를 갖고 있는가?

```
Year_of_Release  Critic_Score  Critic_Count  User_Score  User_Count  Wii  X360  DS  PS  PS2  ...
```

(설명변수의 순서)

	coef	std err	t	P>|t|	[0.025	0.975]
const	93.6995	18.912	4.954	0.000	56.619	130.780
x1	-0.0464	0.009	-4.974	0.000	-0.065	-0.028
x2	0.0062	0.002	3.339	0.001	0.003	0.010
x3	0.0024	0.001	1.601	0.109	-0.001	0.005
x4	-0.0242	0.018	-1.358	0.175	-0.059	0.011
x5	0.0084	0.001	14.861	0.000	0.007	0.010
x6	0.6110	0.147	4.148	0.000	0.322	0.900
x7	0.0654	0.143	0.457	0.647	-0.215	0.346
x8	0.3165	0.153	2.062	0.039	0.016	0.617

⇨ 우선, x1부터 x5까지는 수치형 변수이고, x6부터는 binary형태로 바뀐 명목형 변수이다. 명목형 변수들의 경우 원래의 카테고리인 [Platform, Genre, Publisher, Rating]에 속하는 적어도 하나의 항목이 유의한 p-value를 가지지 않았으므로, 이들은 모두 통계적으로 유의미하지 않는다는 결론을 내릴 수 있다. 그렇다면 남은 x1부터 x5까지의 변수만 고려하면 된다. 이들 중 유의수준 0.01을 만족하는 변수는 [Year_of_Release, Critic_Score, User_Count]이다. 이들은 각각 종속변수와 [음, 양, 양]의 상관관계를 가진다.

[Q8] Test 데이터셋에 대하여 MAE, MAPE, RMSE를 계산하고 그에 대한 해석을 해 보시오.

	RMSE	MAE	MAPE
Game Sales	2.437496e+09	1.164441e+08	9.938059e+07

⇨ Linear Regression에서 하나의 데이터셋의 성능을 평가하는 것은 어려우므로 MAPE로 대강의 성능을 평가할 수 있다. 그런데 여기서 상대 오차는 993.8059e+07%로 지나치게 큰 것을 알 수 있다. 따라서 해당 회귀 모형의 성능이 매우 좋지 않고, 이 데이터셋을 제대로 분석하기 위해서는 선형 회귀 분석이 아닌 다른 방식이 필요하다는 결론을 내릴 수 있다.

[Q9] 만약 원래 변수 수의 절반 이하로 입력 변수를 사용하여 모델을 구축해야 할 경우 어떤 변수들을 선택하겠는가? [Q5]와 [Q7]의 답변을 바탕으로 본인이 선택한 변수들에 대한 근거를 제시하시오.

앞선 분석에 따르면 통계적으로 유의미한 변수들은 [Year_of_Release, Critic_Score, User_Count]이다. 이 때 Critic_Score와 높은 상관관계를 가지는 User_Score, 그리고 User_Count와 높은 상관관계를 가지는 Critic_Count가 빠졌으므로 해당 세 개의 변수가 종속변수를 가장 잘 설명한다고 말할 수 있다. 가장 처음에 정한 설명 변수의 수가 9개였고(Platform, Year_of_Release, Genre, Publisher, Critic_Score, Critic_Count, User_Score, User_Count, Rating), 최종적으로 고른 변수의 수가 3개이므로 적절한 선택이라고 할 수 있다.

[Q10] [Q9]에서 선택한 변수들만을 사용하여 MLR 모형을 다시 학습하고 Adjusted R2, Test 데이터셋에 대한 MAE, MAPE, RMSE를 산출한 뒤, 두 모형(모든 변수 사용 vs. 선택된 변수만 사용)을 비교해 보시오.

```
adr2 = 1-(1-mlr.score(x_trn2,y_trn2))*(x_trn2.shape[0] - 1)/(x_trn2.shape[0] - x_trn2.shape[1]- 1)
adr2
```

```
0.08417117859649736
```

(다시 학습한 모형의 Adjusted R2값)

	RMSE	MAE	MAPE
Game Sales	2.437496e+09	1.164441e+08	9.938059e+07
Game Sales2	6.293527e+01	5.875806e+01	5.063442e+00

⇨ 첫 번째로 수행한 MLR 모형에서 유의미하다고 나온 변수만을 가지고 학습을 시켰으나 놀랍게도 Adjusted R2값이 훨씬 낮아지는 결과가 나왔다. 이러한 결과를 필자는 해당 데이터셋의 근본적인 특성이 선형성과 거리가 지나치게 멀다는 것을 역설적으로 보여주는 것이라 판단했다. RMSE, MAE, MAPE의 성능이 오히려 올랐다는 사실이 이를 뒷받침한다. 앞서 첨부한 coefficient 값 또한 이를 보여준다.

	coef	std err	t	P>|t|	[0.025	0.975]
const	93.6995	18.912	4.954	0.000	56.619	130.780
x1	-0.0464	0.009	-4.974	0.000	-0.065	-0.028
x2	0.0062	0.002	3.339	0.001	0.003	0.010
x3	0.0024	0.001	1.601	0.109	-0.001	0.005
x4	-0.0242	0.018	-1.358	0.175	-0.059	0.011
x5	0.0084	0.001	14.861	0.000	0.007	0.010
x6	0.6110	0.147	4.148	0.000	0.322	0.900
x7	0.0654	0.143	0.457	0.647	-0.215	0.346
x8	0.3165	0.153	2.062	0.039	0.016	0.617

변수의 coefficient보다 상수항의 coefficient값이 지나치게 큰데, 이는 선형성이 정말 작다는 것을 보여준다. 따라서 해당 데이터셋을 분석하기 위해서는 다른 방식의 모델이 필요하며, 정 MLR을 사용하고 싶다면 각 칼럼에 스케일링을 적용하여 feature 간의 단위를 어느정도 맞추는 것이 필요해 보인다. 이는 특정 칼럼에 log를 취하거나 제곱근을 씌우는 등의 방법을 통해 구현이 가능할 것이다.