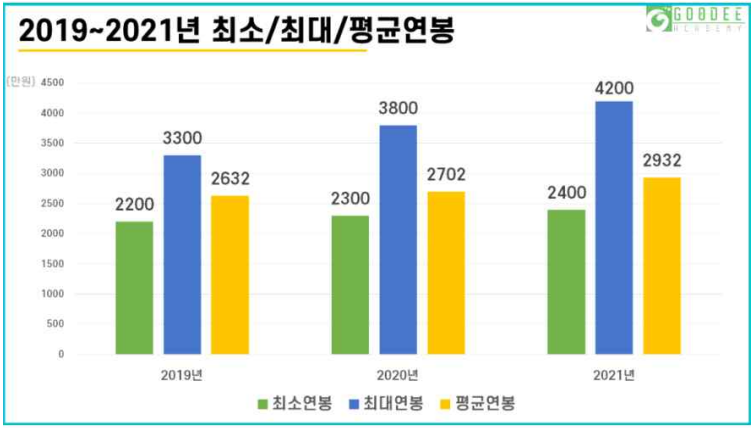
게임산업에서 개발자의 학력 및 업종이 대륙별 수입액의 영향을 미칠까?

-데이터사이언스기초 기말 프로젝트-

빅데이터 20205151 김태호

목차 a table of contents

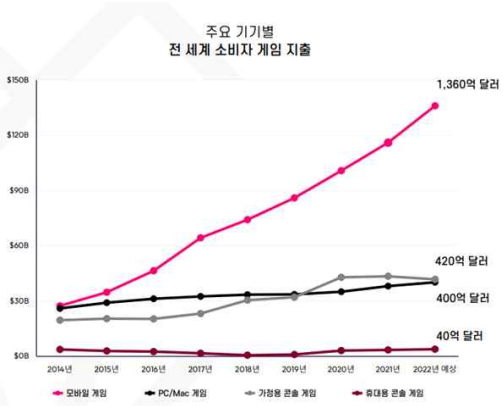
배경 설정



출저-구디아카데미 https://blog.naver.com/goodee0205/222911978475

Part 1

배경 설정



개발자를 꿈꾸고 있어서 개발자의 평균 연봉 및 게임 산업 쪽에도 평소에 관심을 보이고 있었다.

자료를 조사하던 중 왼쪽과 같은 기사를 접했다. 모바일 게임의 게임 지출이 상당히 높은 걸로 보인다.

우리나라도 해외로 게임을 수출할때 왼쪽과 같은 비슷한 그래프를 보이는지, 또는 우리나라에서는 게임으로 해외로 수출했을 때 어떤 분야가 영향을 미쳤는지 궁금해졌다.

게임산업에서 가장 높은 비중을 차지하는 건 어떤 분야인지 궁금하고 게임산업과 학력이 상관관계가 있는지 알아보고 싶었다.

출저-gamevu 김창훈 기자
https://www.gamevu.co.kr/news/articleView.html?idxno=22965

데이터 활용

게임산업_연도별_대륙별_수입.csv

시점	합계	중화권	일본	동남아	북미	유럽	기타	중국
2011	204986	-	163582	0	10522	134	123	30625
2012	179135	-	119397	0	23126	148	166	36298
2013	172229	-	40642	0	35562	20965	25401	49659
2014	165559	-	46774	1387	23136	17631	17470	59161
2015	177492	-	61027	2433	27283	17196	17701	51852
2016	147362	-	51606	4657	18509	8193	3478	60919
2017	262911	67586	110349	7856	52534	22442	2144	-
2018	305781	72850	187226	1626	42207	1733	139	-
2019	298129	132546	47154	192	49737	22364	46134	-
2020	270794	77392	36655	442	139342	10447	6516	-
2021	312331	133072	46365	258	117086	6454	9096	-

출저-kosis 국가 통계포털 게임산업 부분

데이터 활용

게임산업_학력별_업종별_종사자_현황.csv

시점	소계_합계	소계_고졸	소계_초대	소계_대졸	소계_대학	PC게임_합	PC게임_고	PC게임_초	PC게임_대	PC게임_대	모바일게임	모바일게임	모바일게임	모바일게임	모바일게임
2011	51859	2827	8931	38673	1428	122	3	12	107	0	4585	413	895	3214	63
2012	52466	2817	8882	39346	1421	106	2	11	93	0	5823	436	911	4401	75
2013	40541	6145	3082	30612	701	307	16	10	216	64	10215	869	639	8340	366
2014	39221	3942	7613	25411	2255	333	15	99	199	20	9984	963	2095	6488	438
2015	35445	2418	7225	23822	1980	385	24	109	227	25	13106	855	2855	8720	676
2016	33979	2265	8124	20043	3547	212	52	10	115	35	16146	833	3596	9737	1980
2017	34665	2475	9081	21943	1166	13287	1016	3345	8447	479	19686	1213	5255	12557	661
2018	37035	1566	8740	25235	1494	13344	659	3157	8958	570	21742	836	4982	15031	893
2019	39390	1856	6828	28674	2032	13430	847	2510	9296	777	23057	912	3791	17171	1183
2020	44310	1790	8229	32539	1752	14600	1002	3216	9821	561	27028	741	4484	20727	1076
2021	45262	1809	7421	33906	2126	13124	680	2425	9410	609	29015	1012	4369	22199	1435

출저-kosis 국가 통계포털 게임산업 부분

Part 2

데이터 활용

출저-kosis 국가 통계포털 게임산업 부분

게임산업_연도별_직무별_종사자_현황.csv

시점	게임PD	게임기획	웹디자인	그래픽디자	UI 디자인	시스템엔지	게임프로그	웹서비스기	사운드 제	게임운영자	고객지원((	마케팅·홍보	배급 및 유	게임소싱	일반관리직	기타	합계
2009	1301	4423	10928			1778	8933		260	4337	1908	3729			4770	998	43365
2010	1550	5057	11327			2419	10345		340	4502	2618	3799			5458	1170	48585
2011	1765	5092	1968	10322	1425	1937	9233	1733	513	4511	2677	2327	1611	412	5822	511	51859

3개의 csv 파일을 준비했다.

1) 게임산업_연도별_대륙별_수입.csv
2) 게임산업_학력별_업종별_종사자_현황.csv
3) 게임산업_연도별_직무별_종사자_현황.csv

3개의 csv파일을 보면서 먼저 데이터를 파악하는 것이 중요하다. csv파일을 열어서 확인해보니 값이 비어 있는 것 같지만, 숫자로 한 눈에 파악하기는 어렵다고 느껴졌다.

Part 2

데이터 활용

[스크립트]

```
library("tidyverse")
library("dplyr")
library("ggplot2")

#파일 읽기
game <- read_csv("게임산업__학력별_업종별_종사자_현황.csv",locale=locale("ko",encoding="EUC-KR"),col_names=TRUE)

store <- read_csv("대륙별_수입액_현황.csv",locale=locale("ko",encoding="EUC-KR"),col_names=TRUE)

programer <- read_csv("게임산업_직무별_연도별_종사자_현황.csv",locale=locale("ko",encoding="EUC-KR"),col_names=TRUE)
```

[설명]

//파일을 읽기전에

```
library("tidyverse")
library("dplyr")
library("ggplot2")
```

세가지의 library 함수를 먼저 적용했다.

데이터를 가공할 때 필요한 tidyverse와 dplyr 이용할 것이고 그래프를 통해 확인하기 위해 ggplot2 library 함수도 사용할 것이다.

따라서 데이터를 효율적으로 더 편하게 관리하기위해 read_csv 파일로 읽었다.

데이터 활용

[스크립트]

#데이터 확인
str(store)
str(game)
str(programer)

```
모바일게임_대졸 = col_double(),
`모바일게임_대학원졸 이상` = col_double(),
콘솔게임_합계 = col_character(),
콘솔게임_고졸이하 = col_character(),
```

```
합계 = col_double(),
중화권 = col_character(),
일본 = col_double(),
```

#데이터 모든 열을 숫자형으로 변환하기
#char 형태의 열들이 있는데 하나로 통일하는게 데이터를 관리하기 편하다.

```
game <- game %>% mutate_all(as.numeric)
store <- store %>% mutate_all(as.numeric)
programer <- programer %>%
mutate_all(as.numeric)
```

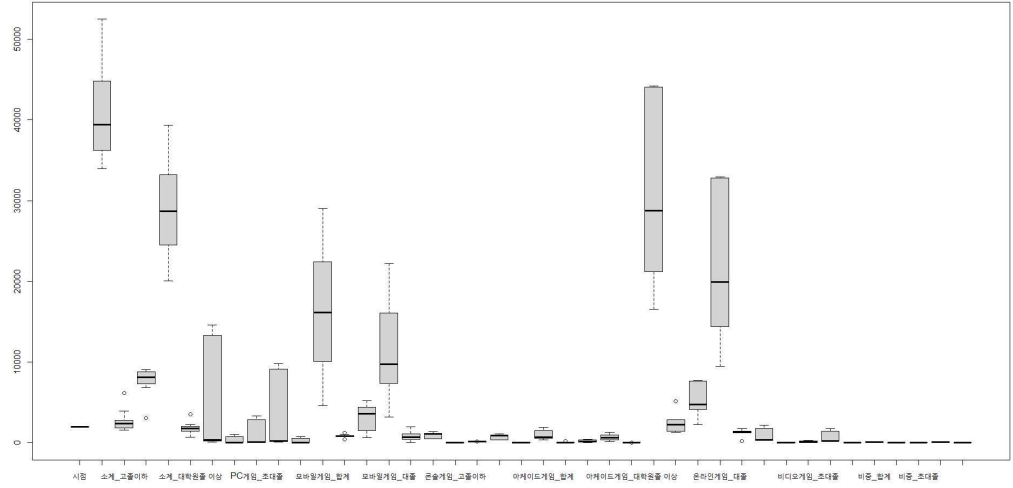
[설명]

//파일을 읽어드리고 데이터가 잘 들어갔는지 또는 데이터의 형태는 어떻게 읽어졌는지 확인 했다.

//데이터를 확인해보니 이처럼 char 또는 double 인 것을 볼 수 있다.

//데이터를 가공하기에 앞서 편하게 할려면 통일을 하는게 맞다고 생각해 모두 num으로 바꿔주었다.

데이터 활용



데이터 활용

[스크립트]

```
#이상값,결측값 확인하기
boxplot(game)
boxplot(store)
boxplot(programer)

#boxplot으로는 이상값과 결측값을 판단하기 어려움.
```

[설명]

/*

boxplot으로 이상값과 결측값이 있는
열을 확인 해봤다.

위에 사진을 보면 확인을 하는 box의 개수가 너무
많다보니 시각적으로 편하게 관측을 할 수 없었고,

다른 데이터또한 마찬가지 였다.

따라서 데이터를 직접 가공하고 그래프를 통해
알아보는게 나아 보았다.

*/

데이터 활용

[스크립트]

```
#programer 데이터 가공
programer
df_long_pro <- programer %>%
  gather(key = "직무별", value = "인원", -시점)

#파이그래프
df_long_pro %>%
  filter(직무별 != "합계") %>%
  ggplot(aes(x = "", y = 인원, fill = 직무별)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  theme_minimal()

#막대그래프
df_long_pro %>%
  filter(직무별 != "합계") %>%
  ggplot(aes(x = 직무별, y = 인원, fill = 직무별)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(legend.position = "none")

#그래프를 통해서 개발자들이 많은 걸 볼 수 있다.
```

```
# A tibble: 51 × 3
   시점 직무별          인원
  <dbl> <chr>          <dbl>
1  2009 게임PD          1301
2  2010 게임PD          1550
3  2011 게임PD          1765
4  2009 게임기획        4423
5  2010 게임기획        5057
6  2011 게임기획        5092
7  2009 웹디자인       10928
8  2010 웹디자인       11327
9  2011 웹디자인        1968
```

[설명]

/*

데이터를 가공하기전에

제목에 맞게 관계를 분석하려면 먼저
위 3개의 데이터가 분석하기에 적합한 데이터
인지 알아봐야 한다.

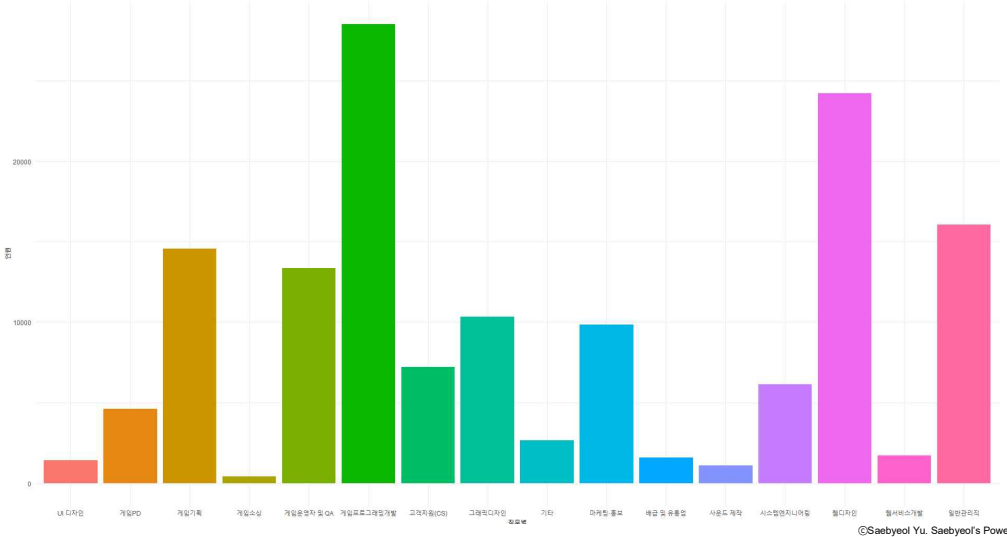
"programer" 데이터프레임에서 "시점" 열을 제외한 모든 열을 "직무별" 열로 정리하고, 해당 열의 값들을 "인원" 열로 통합하여 새로운 데이터프레임인 "df_long_pro"를 생성했다.

그리고 ggplpt 함수를 사용하여 그래프를 생성하고 파이그래프는 x축은 빈 값으로 y축을 인원으로 fill을 직무별로 설정 후 geom_bar를 통해 막대그래프를 그리고 coord_polar 파이그래프로 변환한다. theme_minimal()통해 간단한 스타일로 만든다.

밑 막대그래프도 위에 동일하게 보기 편하게
theme_minimal()과 theme(legend.position = "none")는 간단한 스타일의 테마를 적용하고 범례를 제거한다..
*/

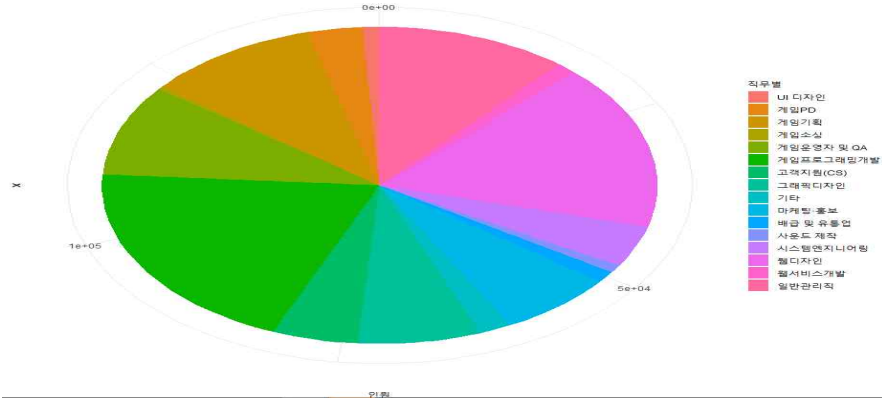
Part 2

데이터 활용



Part 2

데이터 활용



두개의 그래프를 통해서 게임산업에는 코딩을 필요로하는 직무들이 많은 걸 확인 할 수 있었다.

즉 개발자들이 많다고 가정을 하면 제목을
“게임산업에서 개발자의 학력 및 업종이 대륙별 수입액의 영향을 미칠까?”
라는 흥미로운 주제로 선정을 할 수 있다.

데이터 활용

[스크립트]

```
# store 그래프 그리기
#데이터 가공
df_long_store <- store %>%
  gather(key = "나라", value = "수입액", -시점)

 ggplot(df_long_store, aes(x = 시점, y = 수입액, color = 나라, linetype = 나라))
+
   geom_line(size = 1) +
   geom_point(size = 3) +
   labs(x = "시점", y = "수입액") +
   theme_minimal()
 #끊겨있는 그래프를 확인할 수 있음

  #game 그래프 그리기
 #데이터 가공

 game
 df_long_game <- game %>%
   gather(key = "업종_학력", value = "인원", -시점)

 ggplot(df_long_game, aes(x = 시점, y = 인원, color = 업종_학력, linetype = 업
종_학력)) +
   geom_line(size = 1) +
   geom_point(size = 3) +
   labs(x = "시점", y = "인원 수") +
   theme_minimal()+
   facet_wrap(~업종_학력)
```

[설명]

/*

store(대륙별 수입액) game(게임산업_업종_학력)
위 programer처럼 데이터를 편하게 관리하기 위해서 가공을 했다.

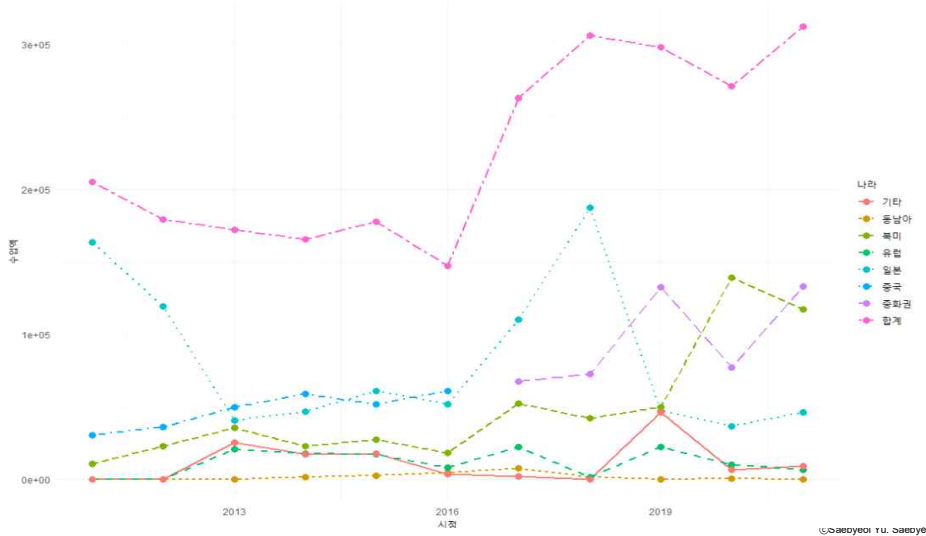
그후 ggplot으로 연도별로 그래프를 확인하였다.
색을 무지개로 주어 보기 구분 하였고

점과 선 그래프를 통해 끊어져있는 변수도 찾을 수 있었다. 즉 결측값이 있다는 건데 그래프를 자세히 보니 비디오게임과 온라인게임이 그래프가 이어져 있지 않았다.

비디오게임은 콘솔게임 그래프와 비교했을때 비디오게임은 더이상 통계하지않고 콘솔을 통계함으로써 둘이 합쳐진 것을 예측할 수 있었고, 이를 통해 온라인게임또한 시대가 바뀌면서 통계를 안한 다는 것을 알게 되었다. 이유는 게임에서 온라인 오프라인을 구분 짓는게 모호 해졌다고 생각한다.

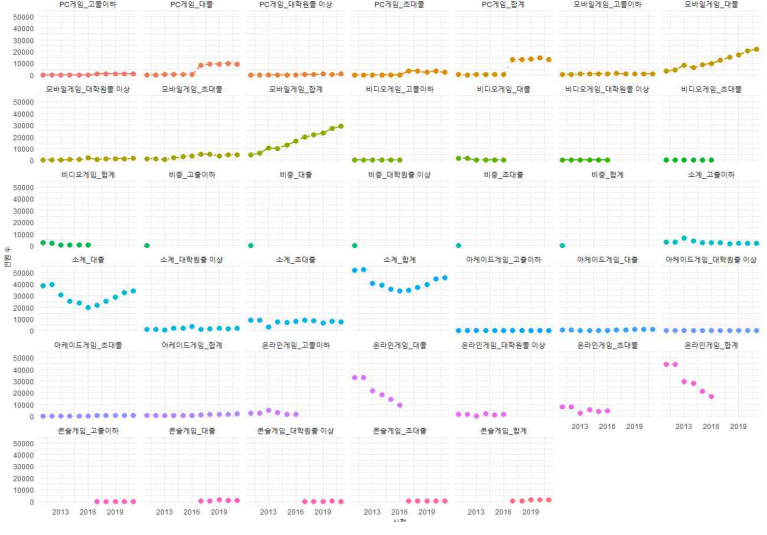
*/

데이터 활용



Part 2

데이터 활용

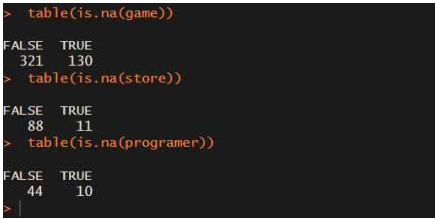


Part 2

데이터 활용

[스크립트]

```
table(is.na(game))
table(is.na(store))
table(is.na(programer))
```



[설명]

/*

NA값이 들어있는 건 그래프와 is.na()를 통해 알 수 있었다..

그럼 0으로 되어 있는 값들을 그대로 둬야하나 NA로 치환을 해야하나 고민했다. 물론 0과 NA는 엄연히 다른 값이다.

만약 그래프에서 눈에 띄게 갑자기 그래프가 올라가는 그래프가 있었다면 NA로 바꿔야하겠지만,

하지만 데이터를 확인해보니 0이 NA처럼 결측값이 아닌 실제로 0인 수치인 걸 그래프의 선으로 알 수 있었다.

*/

데이터 활용

[스크립트]

```
game%>%
  slice(1:11) %>%
  select(contains("합계")) %>%
  summarise_all(sum, na.rm = TRUE) %>%
  gather(key = "열", value = "합계") %>%
  arrange(desc(합계))
```

[설명]

/*

dplyr함수를 사용하여 game 데이터프레임의 11개행을 선택, 그 중에 "합계"를 포함하는 열을 선택하고 선택된 모든 값을 합산한다. 이때 NA는 제외. 열을 행으로 변환하여 열과 합계 열을 만든다. arrange를 통해 내림차순으로 정렬한다.

*/

```
# A tibble: 8 × 2
  열                  합계
  <chr>              <dbl>
1 소계_합계          454173
2 온라인게임_합계    183447
3 모바일게임_합계    180387
4 PC게임_합계         69250
5 아케이드게임_합계   11095
6 비디오게임_합계      5326
7 콘솔게임_합계        4668
8 비중_합계             100
```

종사하는 인원수가 많을수록 대륙별_수입액과의 유의미한 관계인지 확인도 할 수 있게 순위를 파악한다.

데이터 활용

[스크립트]

```
First_Data <- data.frame(c(store[,1],store[,2],game[,2],
game[,27],game[,12], game[,7]),game[,22])

#헷갈리지 않게 열 이름 변경
names(First_Data)[2] <- "수입액"
First_Data
```

```
   시점 수입액 소계_합계 온라인게임_합계
1  2011 204986       51859          44221
2  2012 179135       52466          44036
3  2013 172229       40541          29267
4  2014 165559       39221          28202
5  2015 177492       35445          21198
6  2016 147362       33979          16523
7  2017 262911       34665             NA
8  2018 305781       37035             NA
9  2019 298129       39390             NA
10 2020 270794       44310             NA
11 2021 312331       45262             NA
   모바일게임_합계 PC게임_합계 아케이드게임_합계
1             4585         122              713
2             5823         106              719
3            10215         307              430
4             9984         333              395
5            13106         385              441
6            16146         212              716
7            19686       13287             1215
8            21742       13344             1420
9            23057       13430             1555
```

[설명]

모델을 만들고 단순 선형회귀 분석을 하기 전에
내가 원하는 데이터를 뽑아 새로운 데이터프레임을 만들어준다.

또한 이름을 헷갈리지 않게 하기위해 names()를 통해 이름을 변경한다.

가설설정을 하자면

H0: b = 0 (Y는 X의 함수가 아니다, X와 Y는 회귀 관계가 성립되지 않는다.)
H1: b ≠0 (Y는 X의 함수이다, X와 Y는 회귀 관계가 성립된다.)

X=OO게임의 합계 Y=수입액이 된다.
따라서 총 4가지의 경우를 확인할 것이다.
(온라인,pc,모바일,아케이드)

Part 2

데이터 활용

[스크립트]

```
#모델 만들고 분석 단순 선형회귀 분석 결과
#첫번째 온라인
First_model_1 = lm(수입액~온라인게임_합계,data=First_Data)
plot(First_model_1,col="blue")
abline(First_model_1)

summary(First_model_1)

#두번째 모바일
First_model_2 = lm(수입액~모바일게임_합계,data=First_Data)
plot(First_model_2,col="blue")
abline(First_model_2)

summary(First_model_2)

#세번째 pc게임
First_model_3 = lm(수입액~PC게임_합계,data=First_Data)
plot(First_model_3,col="blue")
abline(First_model_3)

summary(First_model_3)

#네번째 아케이드게임

First_model_4 = lm(수입액~아케이드게임_합계,data=First_Data)
plot(First_model_4,col="blue")
abline(First_model_4)

summary(First_model_4)
```

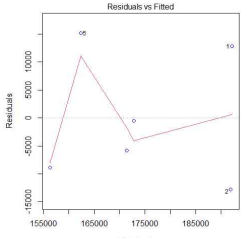
[설명]

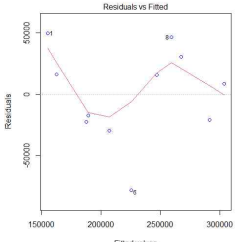
lm()를통해 온라인 게임과 수입액과의 관계를 모델링을 하였고,시각적으로 판단할 수 있게 plot과 abline 함수를 이용해 회귀 모델이 산점도와 추세선을 그래프로 표현하였다.

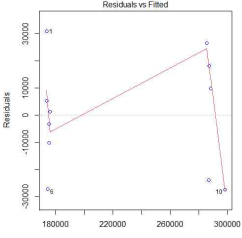
이와 동일한 방법으로 모바일게임,pc게임,아케이드게임도 만들었고

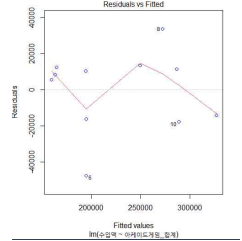
마지막 sumary()함수로 회귀모델의 요약 통계 및 유의성 검정결과를 출력했다.

데이터 활용









4개의 그래프 중에서 3번째의 있는 pc게임의 그래프가

Residuals vs Fitted Plot을 통해 잔차가 예측값 주위에 무작위로 분포되어 있는지 확인했을 때 첫 번째(온라인게임)그래프를 제외하고 점들이 규칙적으로 분포되있는게 보이고 3개의 그래프의 모양이 비슷한 걸 알 수 있다.

다음 summary로 요약 통계량을 통해 정확하게 분석 결과를 확인해 보았다.

유의 수준을 0.05 설정하고 p-value가 유의 수준 보다 작을 경우, 우연히 발생한 것이 아닌 실제로 효과가 있다는 의미이다. 그리고 R-squared는 종속 변수의 변동성과 독립변수들이 얼마나 잘 설명하는지를 나타내는 지표이다. 따라서 1에 가까울수록 더 잘 설명한다는 의미이다.

데이터 활용

[콘솔 창]

```
Residuals:
       1        2        3        4        5
 12832.7 -12778.5   -536.3  -5825.5  15188.2
       6
 -8880.6

Coefficients:
                 Estimate Std. Error t value
(Intercept)      1.348e+05  1.630e+04   8.270
온라인게임_합계 1.296e+00  5.044e-01   2.571
                Pr(>|t|)
(Intercept)      0.00117 **
온라인게임_합계  0.06194 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12960 on 4 degrees of freedo
m
  (결측으로 인하여 5개의 관측치가 삭제되었습니다.)
Multiple R-squared:  0.6229,    Adjusted R-squared:
0.5286
F-statistic: 6.608 on 1 and 4 DF,  p-value: 0.06194
```

[설명]

첫 번째 모델에서 p-value는 0.06194로, 0.05보다 큰 값이므로 해당 모델에서는 온라인 게임의 합계가 수입액에 통계적으로 유의미한 영향을 미치지 않고

Multiple R-squared는 0.6229 애매한 수치를 표현하여

첫 번째 모델은 수입액과의 관계에서 유의미한 관계는 가지지 않는다는 걸 알 수 있다.

데이터 활용

[콘솔 창]

```
Call:
lm(formula = 수입액 ~ 모바일게임_합계, data = First_Dat
a)

Residuals:
   Min     1Q Median     3Q    Max
-78076 -21596   8727  23556  49770

Coefficients:
                Estimate Std. Error t value
(Intercept)     1.274e+05  2.757e+04   4.619
모바일게임_합계 6.074e+00  1.513e+00   4.016
                Pr(>|t|)
(Intercept)      0.00126 **
모바일게임_합계  0.00304 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39950 on 9 degrees of freedo
m
Multiple R-squared:  0.6418,    Adjusted R-squared:
0.602
F-statistic: 16.13 on 1 and 9 DF,  p-value: 0.003038
```

[설명]

두 번째 모델에서 p-value는 0.00304로, 0.05보다 작은 값이므로 해당 모델에서는 모바일 게임의 합계가 수입액에 통계적으로 유의미한 영향을 미치고 있고,

Multiple R-squared는 0.6418 애매한 수치를 표현하여

두번째 모델은 수입액과의 관계에서 p-value는 좋은 수치지만 적합한 유의미한 관계는 가지지 않는다는 걸 알 수 있다.

데이터 활용

[스크립트]

```
Call:
lm(formula = 수입액 ~ PC게임_합계, data = First_Data)

Residuals:
   Min     1Q Median     3Q    Max
-27500 -17158   1278  14079  31031

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.729e+05  9.124e+03  18.951 1.46e-08
PC게임_합계 8.588e+00  9.972e-01   8.612 1.22e-05

(Intercept) ***
PC게임_합계 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21960 on 9 degrees of freedo
m
Multiple R-squared:  0.8918,    Adjusted R-squared:
0.8798
F-statistic: 74.16 on 1 and 9 DF,  p-value: 1.223e-05
```

[설명]

세 번째 모델에서 p-value는 1.223e-05로, 0.05보다 작은 값이므로 해당 모델에서는 PC 게임의 합계가 수입액에 통계적으로 유의미한 영향을 미치고 있고,

Multiple R-squared는 0.8918 높은 수치로
4개의 모델 중에서 가장 높은 수치이다.

세 번째 모델은 수입액과의 관계에서 p-value는 아케이드 게임모델에 비해서 낮지만 R-squared가 높아 좋은 결과를 보이고 있다.

데이터 활용

[스크립트]

```
Call:
lm(formula = 수입액 ~ 아케이드게임_합계, data = First_Dat
a)

Residuals:
   Min     1Q Median     3Q    Max
-47616 -15174   8521  12000  33831

Coefficients:
                 Estimate Std. Error t value
(Intercept)      116693.51   15361.28   7.597
아케이드게임_합계    109.34      13.55   8.069
                 Pr(>|t|)
(Intercept)      3.34e-05 ***
아케이드게임_합계 2.07e-05 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23260 on 9 degrees of freedo
m
Multiple R-squared:  0.8786,    Adjusted R-squared:
0.8651
F-statistic: 65.11 on 1 and 9 DF,  p-value: 2.066e-05
```

[설명]

세 번째 모델에서 p-value는 2.066e-05로, 0.05보다 작은 값이므로 해당 모델에서는 아케이드 게임의 합계가 수입액에 통계적으로 유의미한 영향을 미치고 있고,

Multiple R-squared는 0.8786 높은 수치로
pc게임의 모델 보다는 낮지만 높은 수치다.

네 번째 모델은 수입액과의 관계에서 R-squared가 가장 좋은 수치는 아니지만 p-value는 4개의 그래프중 가장 좋은 결과를 보이고 있다.

데이터 활용

따라서 3개의 분석 결과를 알 수 있다. summary()로 보기전에 그래프를 통해 봤을때 첫 번째 그래프 빼고 나머지 그래프들이 비슷한 모양을 보이고 있었고, 3개가 유의미한 결과를 가지고 있다는 걸 알 수 있었다.

첫 번째 종사자 합계 순위를 정렬 했을때 4개의 모델 중 온라인게임이 1순위로 높았고 아케이드 게임이 마지막이었다 하지만 통계를 직접 보니 아케이드 게임의 모델이 수입액과의 더 적합한 모델이였다.

귀무가설 : 종사자의 합계는 대륙별 수입액과 관련이 없다.

대립 가설 : 종사자의 합계는 대륙별 수입액과 관련이 있다

따라서 귀무가설이 받아 들여진다.

두 번째로 학력별_업종별이 대륙별_수입액과 관련이 있을까? 라는 분석을 들어가기전에 내가 관심있는 분야인 pc게임과 모바엘 게임 중에서 pc게임의 모델이 수입액과의 관계에서 가장 적합한 모델이었다. 따라서 pc게임산업의 다양한 학력또한 영향이 있는지 분석을 할 수 있는 명분이 생겼다.

데이터 활용

[스크립트]

First_model_multi = lm(수입액 ~ 아케이드게임_합계 + 모바일게임_합계 + PC게임_합계, data = First_Data)

summary(First_model_multi)

```
Call:
lm(formula = 수입액 ~ 아케이드게임_합계 + 모바일게임_합계 +
    PC게임_합계, data = First_Data)

Residuals:
   Min     1Q Median     3Q    Max
-20856 -15361   5238   9787  19170

Coefficients:
                 Estimate Std. Error t value
(Intercept)      160771.117  19997.552   8.040
아케이드게임_합계     68.166     29.972   2.274
모바일게임_합계       -2.647      1.606  -1.648
PC게임_합계            6.489      2.407   2.696
                Pr(>|t|)
(Intercept)     8.83e-05 ***
아케이드게임_합계  0.0571 .
모바일게임_합계   0.1434
PC게임_합계       0.0308 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18240 on 7 degrees of freedo
m
Multiple R-squared:  0.9419,    Adjusted R-squared:
0.917
F-statistic: 37.83 on 3 and 7 DF,  p-value: 0.0001075
```

[설명]

분석을 진행하기에 앞서 정말 pc게임이 다양한 업종에서 통계적으로 유의미한 변수를 판단하기 위해 다중 선형 회귀 분석을 하였다.

유의미한 결과를 보여주는 세개의 모델에서

학력의 수준이 게임산업_대륙별_수입액 관계가 있을까?

를 알아보기 위해 가장 적합한 업종의 설명변수를 찾고 싶었다.

다중 선형 회귀 분석 결과를 통해 수입액과의 관계에서 pc게임의_합계가 p-value가 0.03으로 0.05 값보다 작아서 통계적으로 유의미한 변수로 판단되어 pc게임의 합계를 설명변수로 채택하였다.

데이터 활용

[스크립트]

```
Second_Data <-
data.frame(c(store[,1],store[,2],game[,7],game[,8],game[,
9],game[,10],game[,11]))
names(Second_Data)[2] <- "수입액"
```

[설명]

#가장 유의미한 결과를 보여준 Pc게임에서 학력수준은 상관관계가 있을까?를 알아보기위해

위와 같은 방법으로 두 번째 모델을 만들기 전 새로운 데이터프레임을 만들었다.

```
   시점  수입액 PC게임_합계 PC게임_고졸이하
1  2011 204986         122              3
2  2012 179135         106              2
3  2013 172229         307             16
4  2014 165559         333             15
5  2015 177492         385             24
6  2016 147362         212             52
7  2017 262911       13287           1016
8  2018 305781       13344            659
9  2019 298129       13430            847
10 2020 270794       14600           1002
11 2021 312331       13124            680
   PC게임_초대졸 PC게임_대졸 PC게임_대학원졸.이상
1             12         107                    0
2             11          93                    0
3             10         216                   64
4             99         199                   20
5            109         227                   25
6             10         115                   35
7           3345        8447                  479
8           3157        8958                  570
9           2510        9296                  777
10          3216        9821                  561
11          2425        9410                  609
```

데이터 활용

[스크립트]

```
rapid_Model = function(x){

   Second_Model<- lm(수입액 ~ Second_Data[,x+2],
data = Second_Data)
   return(summary(Second_Model))
}

#순서 PC게임_합계 PC게임_고졸이하 PC게임_초대졸
PC게임_대졸 PC게임_대학원졸 이상으로 분류 함.

i=1
while(i<=5){
 print(paste(i,"번째"))
  print(rapid_Model(i))

  i=i+1
}
```

[설명]

함수를 만들어서 모델을 일일이 만들고 summary를 반복적으로 사용하지 않게 하기 위해 코드를 만들었고

while 반복문을 통해 5번 출력하였다.

#PC게임_합계 PC게임_고졸이하 PC게임_초대졸 PC게임_대졸 PC게임_대학원졸 순서이다.

데이터 활용

```
[1] "1 번째"

Call:
lm(formula = 수입액 ~ Second_Data[, x + 2], data = Sec
ond_Data)

Residuals:
   Min     1Q Median     3Q    Max
-27500 -17158   1278  14079  31031

Coefficients:
                      Estimate Std. Error t value
(Intercept)          1.729e+05  9.124e+03  18.951
Second_Data[, x + 2] 8.588e+00  9.972e-01   8.612
                     Pr(>|t|)
(Intercept)          1.46e-08 ***
Second_Data[, x + 2] 1.22e-05 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21960 on 9 degrees of freedo
m
Multiple R-squared:  0.8918,    Adjusted R-squared:
0.8798
F-statistic: 74.16 on 1 and 9 DF,  p-value: 1.223e-05
```

1번째 모델 : pc게임_합계

Multiple R-squared: 0.8918

p-value: 1.223e-05

데이터 활용

```
[1] "2 번째"

Call:
lm(formula = 수입액 ~ Second_Data[, x + 2], data = Sec
ond_Data)

Residuals:
   Min     1Q Median     3Q    Max
-41475 -23213  -3756  20533  49653

Coefficients:
                      Estimate Std. Error t value
(Intercept)         178269.30   13643.16  13.067
Second_Data[, x + 2]    124.13      23.67   5.244
                    Pr(>|t|)
(Intercept)         3.72e-07 ***
Second_Data[, x + 2] 0.000532 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33150 on 9 degrees of freedo
m
Multiple R-squared:  0.7534,    Adjusted R-squared:
0.726
F-statistic:  27.5 on 1 and 9 DF,  p-value: 0.000532
```

2번째 모델 : pc게임_고졸

Multiple R-squared: 0.7534

p-value: 0.000532

데이터 활용

```
[1] "3 번째"

Call:
lm(formula = 수입액 ~ Second_Data[, x + 2], data = Sec
ond_Data)

Residuals:
   Min     1Q Median     3Q    Max
-38723 -20149  -2740  19510  45212

Coefficients:
                      Estimate Std. Error t value
(Intercept)          1.761e+05  1.148e+04  15.350
Second_Data[, x + 2] 3.752e+01  5.757e+00   6.517
                     Pr(>|t|)
(Intercept)          9.23e-08 ***
Second_Data[, x + 2] 0.000109 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27910 on 9 degrees of freedo
m
Multiple R-squared:  0.8251,    Adjusted R-squared:
0.8057
F-statistic: 42.47 on 1 and 9 DF,  p-value: 0.0001093
```

2번째 모델 : pc게임_초대졸

Multiple R-squared: 0.8251

p-value:0.0001093

데이터 활용

```
[1] "4 번째"

Call:
lm(formula = 수입액 ~ Second_Data[, x + 2], data = Sec
ond_Data)

Residuals:
   Min     1Q Median     3Q    Max
-27200 -13510   2082  12829  31110

Coefficients:
                      Estimate Std. Error t value
(Intercept)          1.725e+05  8.348e+03  20.664
Second_Data[, x + 2] 1.278e+01  1.346e+00   9.493
                     Pr(>|t|)
(Intercept)          6.81e-09 ***
Second_Data[, x + 2] 5.51e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20110 on 9 degrees of freedo
m
Multiple R-squared:  0.9092,    Adjusted R-squared:
0.8991
F-statistic: 90.12 on 1 and 9 DF,  p-value: 5.508e-06
```

4번째 모델 : pc게임_대졸

Multiple R-squared: 0.9092

p-value: 5.508e-06

데이터 활용

```
[1] "5 번째"

Call:
lm(formula = 수입액 ~ Second_Data[, x + 2], data = Sec
ond_Data)

Residuals:
   Min     1Q Median     3Q    Max
-30966 -10794  -1654  15060  33455

Coefficients:
                      Estimate Std. Error t value
(Intercept)          171530.51    8986.29  19.088
Second_Data[, x + 2]    194.23      21.91   8.864
                     Pr(>|t|)
(Intercept)          1.37e-08 ***
Second_Data[, x + 2] 9.67e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21400 on 9 degrees of freedo
m
Multiple R-squared:  0.8972,    Adjusted R-squared:
0.8858
F-statistic: 78.57 on 1 and 9 DF,  p-value: 9.671e-06
```

4번째 모델 : pc게임_대학원졸

Multiple R-squared: 0.8972

p-value: 9.671e-06

데이터 활용

5개의 분석 중 첫 번째 모델 : pc게임_합계를 제외하고 나머지 모델의 분석 결과를 알 수 있다.

귀무가설 : 학력의 수준이 높을 수록 대륙별 수입액과 관련이 없다.

대립 가설 : 학력의 수준이 높을 수록 대륙별 수입액 관련이 있다.

4번째 모델(pc게임_대졸)이 p-value와 Multiple R-squared를 확인 했을때 p-value: 5.508e-06 으로 4개의 모델 중 제일 작은 값으로 유의미한 값을 나타내고 있었고, Multiple R-squared: 0.9092은 제일 높게 이 분석의 설명력을 나타내고 있었다. 그리고 5번째 모델(pc게임_대학원졸)도 4번째 모델과 차이가 없는 걸로 보아 학력이 높을 수록 유의미한 결과를 가진다. 반면에 2번째(pc게임_고졸)와 3번째(pc게임_초대졸)은 두 모델에 비해 p-value는 높고 R-squared가 낮은결과를 가지고 있다.

따라서 학력의 수준이 높을수록 대륙별 수입액과 관련이 있다.라는 대립가설이 채택되어 진다.

데이터 활용

[스크립트]

```
Second_Data
Second_Model<- lm(수입액 ~ PC게임_대졸, data =
Second_Data)
summary(Second_Model)

new_data <- data.frame(PC게임_대졸 = 17805)
predicted_income <- predict(Second_Model, newdata =
new_data)
predicted_income
```

	시점	수입액	PC
1	2011	204986	
2	2012	179135	
3	2013	172229	
4	2014	165559	
5	2015	177492	
6	2016	147362	
7	2017	262911	
8	2018	305781	
9	2019	298129	
10	2020	270794	
11	2021	312331	

```
> new_data <- data.frame(PC게임_대졸 =17804)
> predicted_income <- predict(Second_Model, newdata =
new_data)
> predicted_income
       1
399994.1
> #예측! 성공!
> new_data <- data.frame(PC게임_대졸 =17805)
> predicted_income <- predict(Second_Model, newdata =
new_data)
> predicted_income
       1
400006.9
```

[설명]

Second_Data를 보면 수입액이 점차 좋아지기는 하지만 수입액이 400000이 넘은 통계가 없다. 그래서 회귀식을 통해 알아보고 싶었다.

이전에 관계를 검정 할때 pc게임_대졸이 가장 적합한 모델이어서 이 모델로 수입액을 예측 해봤다.

회귀식
수입액 = 172,500 + 12.78 * PC게임_대졸

이를 통해서 predict() 함수를 사용해 Second_Model을 기반으로 new_data의 수입액을 예측한다.

따라서 (pc게임_대졸의 수>17805) 이면
수입액이400000 넘어가는 걸 알 수 있었다.

Part 3

마무리

“게임산업에서 개발자의 학력 및 업종이 대륙별 수입액의 영향을 미칠까?”

라는 분석을 하기 전에 게임 산업에 개발자가 많이 있는지, 적절한 데이터를 사용하는 것인지 알기위해
데이터를 가공하여 그래프로 시각화하여 적합한지 판단하였고,

판단 후 게임산업_학력별_업종별_현황.csv 과 대륙별_수입액_현황.csv 파일도 그래프 보여주어 이를 통해
결측값,이상값을 확인하였다.

내가 원하는 가설에 도달하기전에 어떤 업종이 대륙별 수입액에 영향을 끼쳤는지 알아보기 위해
다양한 업종들을 모델링하여 선형 회귀 분석을 통해 pc_게임의 업종이 대륙별 수입액에 영향을 끼치는 걸 알았다.
또한 종사자 수가 많을 수록 영향을 끼치지 않는다는 사실도 통계를 통해 알 수 있었다.

pc게임산업의 다양한 학력으로 학력 수준이 높을 수록 수입액에 영향을 끼칠까? 라는 가설을 세워
위와 동일한 방법으로 선형 회귀 분석을 통해 학력이 높을 수록 수입액에 영향을 끼친다는 것도 통계를 통해
확인할 수 있었다.

마지막으로 가장 적한한 모델의 학력을 선정하여 수입액이 목표액을 도달하려면 얼마나 더 많은 사람이 필요한지 예측을 통해 알 수 있었다.

따라서 개발자의 학력은 수준이 높을수록 대륙별 수입액의 영향을 미친다.는 것을 데이터 분석을 알 수 있었다.