BigData Procesing

PROF Aziz Nasridinov

Introduction

Lee u seog
Presentation
regression analysis
data collection

Park Ji su
Ppt production
Decision tree

Hwang Se young
Knn
Data Processing

Jang hee ju
Data Processing
Basic StatisticalAnalysis

Movie

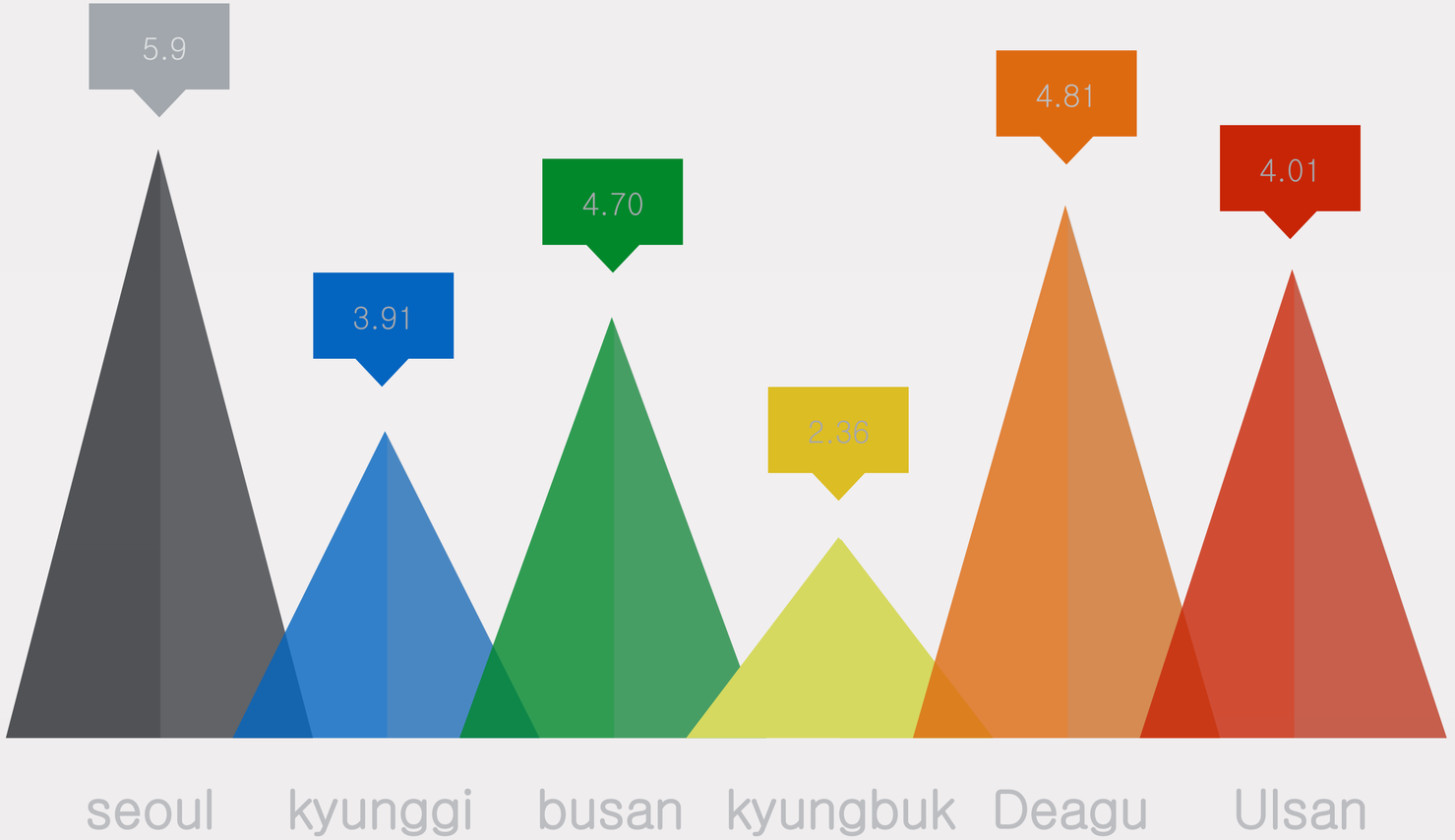
범죄도시

Until This year 19 movies have reached 10 million.
The number of movie audiences increases as the level of people's cultural life increases.
We analyzed the movie data to predict the number of audiences in the movie.

Number of visits person by region in 2016

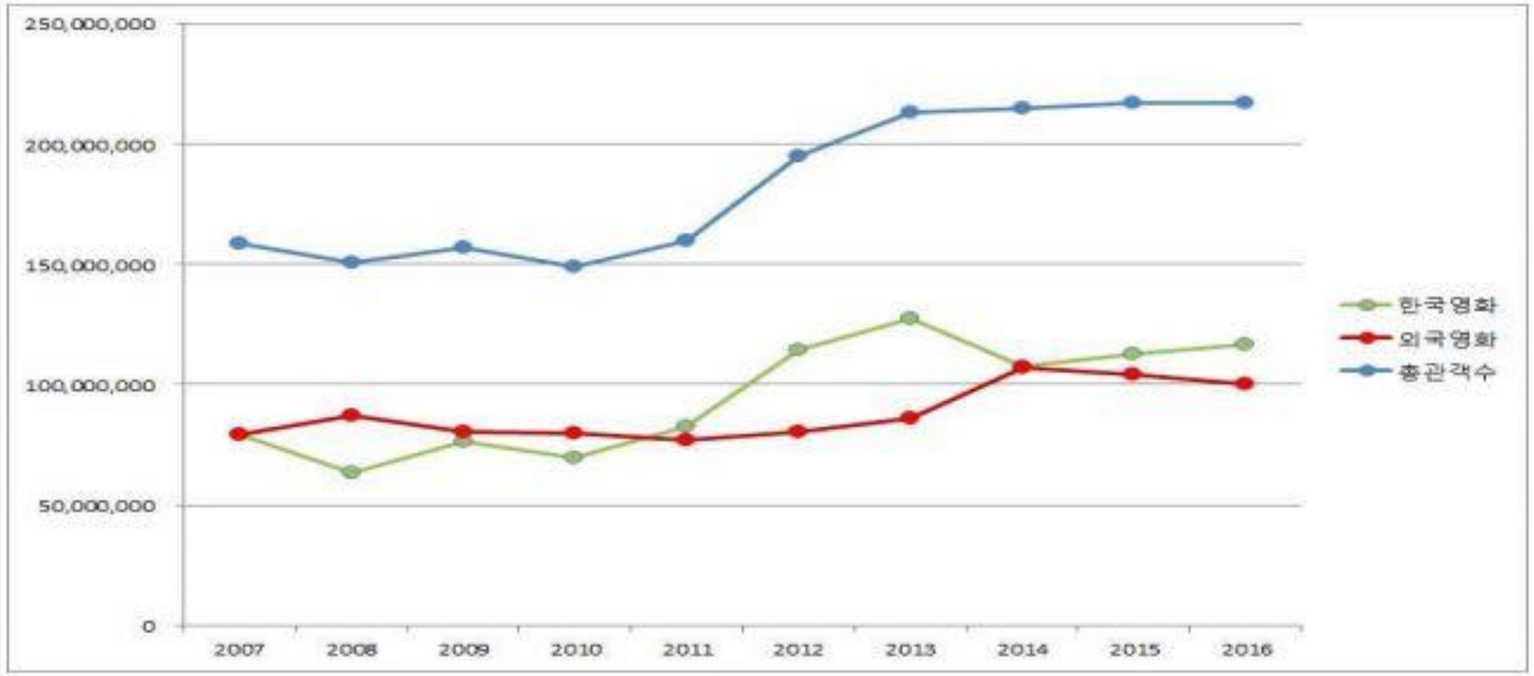
Information

Population by area group Number of visitors per person Station The city of Seoul was the highest at 5.90. Gwangju Metropolitan City recorded 5.40 times, the second highest viewing Respectively.



출처 : 한국문화산업 진흥회

In 2016, the total number of audiences in theaters totaled 211.72 million, down 0.1% from the previous year. Since 2011 By 2015, the total number of audiences in the theater has increased for the fifth consecutive year. . The number of Korean movie audiences rose to 111.55 million, up 3.2% from the previous year. With this Korean films have surpassed 100 million audiences for five consecutive years since 2012.



Korean Film Industry Closing in 2016

Analysis sequence

Step 1.

Information of Data set

Step 2.

Basic statistical analysis

Step 3.

knn

Step 4.

Decision tree

Step 5.

Regression analysis

Step 6.

apply

STEP 1.
INFORMATION OF DATA

STEP 1.
Movie Data Set

Initial data set, Source: Competition

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	X1st_pe	series_N	series_Y	TV	Korea	US	Jepan	12up	15up	19up	distributor	action	comedy	drama	horror	sf	mello	animation	hiswar	crime	X1st_goldti	X2nd_gold	naver_goo	naver_bad	naver_pe	blog	actor	director
2	1344283	0	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	0	6919	866	2339.34	199	0	0
3	1196692	1	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1	3751	383	1470.5	158	0.0238	0
4	2009291	0	1	0	0	1	0	1	0	0	1	1	0	0	0	1	0	0	0	0	3	3	7337	677	5093.34	245	0.0938	0
5	415520	1	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	12208	1071	3652.4	382	0.1333	0
6	694266	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	1	1	4570	2003	4794.87	227	0.0588	0.067
7	610123	1	0	0	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	2439	350	1038.86	76	0	0
8	1192642	1	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	21361	4102	12298.51	380	0.08	0.333
9	5469383	1	0	0	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	1	4	4	46891	2492	6706.54	230	0.3333	0
10	2183079	1	0	0	1	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	15918	2041	4407.66	156	0.3227	0
11	1189551	1	0	0	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	5068	759	8507.7	60	0	0
12	320108	1	0	0	1	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	8640	1340	1991.06	179	0.1103	0
13	287965	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	6385	658	4189.34	136	0.0238	0
14	373983	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	3	3583	625	3952.55	281	0.0526	0.278
15	335848	1	0	0	0	1	0	1	0	0	1	1	1	0	0	1	0	0	0	0	0	0	2374	328	2752.68	83	0.1364	0
16	3024092	1	0	0	1	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	1	1	15570	2644	7205.28	470	0.2788	0.25
17	632425	0	0	1	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	4457	1042	7299.88	237	0.0862	0
18	1428021	1	0	0	1	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	9879	1277	5238.33	341	0.2033	0
19	237196	1	0	0	0	1	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	2541	363	6660.5	125	0.0882	0

01 X1st_pe

1st Cumulative number of movie theaters

02 Series_N & Y

Whether or not there is existence of series.

03 TV

TV simultaneous screening

04 Korea , US, Japen

Korea, United States, Japan Which country's movies. Other is 0

05 12,15,19 up, distributor

Age limit of movies Whether or not there is a distributor

06 Genre

Various genres such as action, horror, comedy,etx..

07 X1st_goldtimes ,2nd

How much overlap with Golden Holidays

08 naver_good,bad ,pe

Good, bad evaluation from Naver.
Multiply with Movie audience and Rating

09 Blog,actor,director

Number of posting
The influence of actors and directors

BIG DATA

Movie Data Set

```
read.csv("movie_dataset - 복사본.csv", stringsAsFactors =
e)
'e':   328 obs. of  29 variables:
       : int  1344283 1196692 2009291 415520 694266 6101
N      : int  0 1 0 1 1 1 1 1 1 1 ...
Y      : int  1 0 1 0 0 0 0 0 0 0 ...
       : int  0 0 0 0 0 0 0 0 0 0 ...
       : int  1726416 1573959 2516537 504146 948493 8234
       : int  0 0 0 1 1 0 1 1 1 0 ...
       : int  0 1 1 0 0 1 0 0 0 1 ...
       : int  0 0 0 0 0 0 0 0 0 0 ...
       : int  0 0 1 1 0 1 0 0 0 0 ...
       : int  1 1 0 0 0 0 0 1 1 1 ...
       : int  0 0 0 0 1 0 1 0 0 0 ...
utor   : int  0 0 1 1 1 0 0 1 1 0 ...
       : int  1 0 1 0 0 1 1 0 0 0 ...
       : int  0 0 0 0 0 0 0 1 0 0 ...
       : int  0 0 0 1 0 0 1 0 1 1 ...
       : int  1 1 0 0 0 0 0 0 0 1 ...
       : int  0 0 1 1 0 1 0 0 0 0 ...
       : int  0 0 0 0 0 0 0 0 0 0 ...
on     : int  0 0 0 0 0 0 0 0 0 0 ...
       : int  0 0 0 0 1 0 0 0 0 0 ...
```

str(movie)

```
> summary(movie)
     X1st_pe
 Min.   :  25506
 1st Qu.: 173719
 Median : 352570
 Mean   : 737607
 3rd Qu.:1000110
 Max.   :6236450
```

summary(movie)

Step 2.

Basic Statistical Analysis

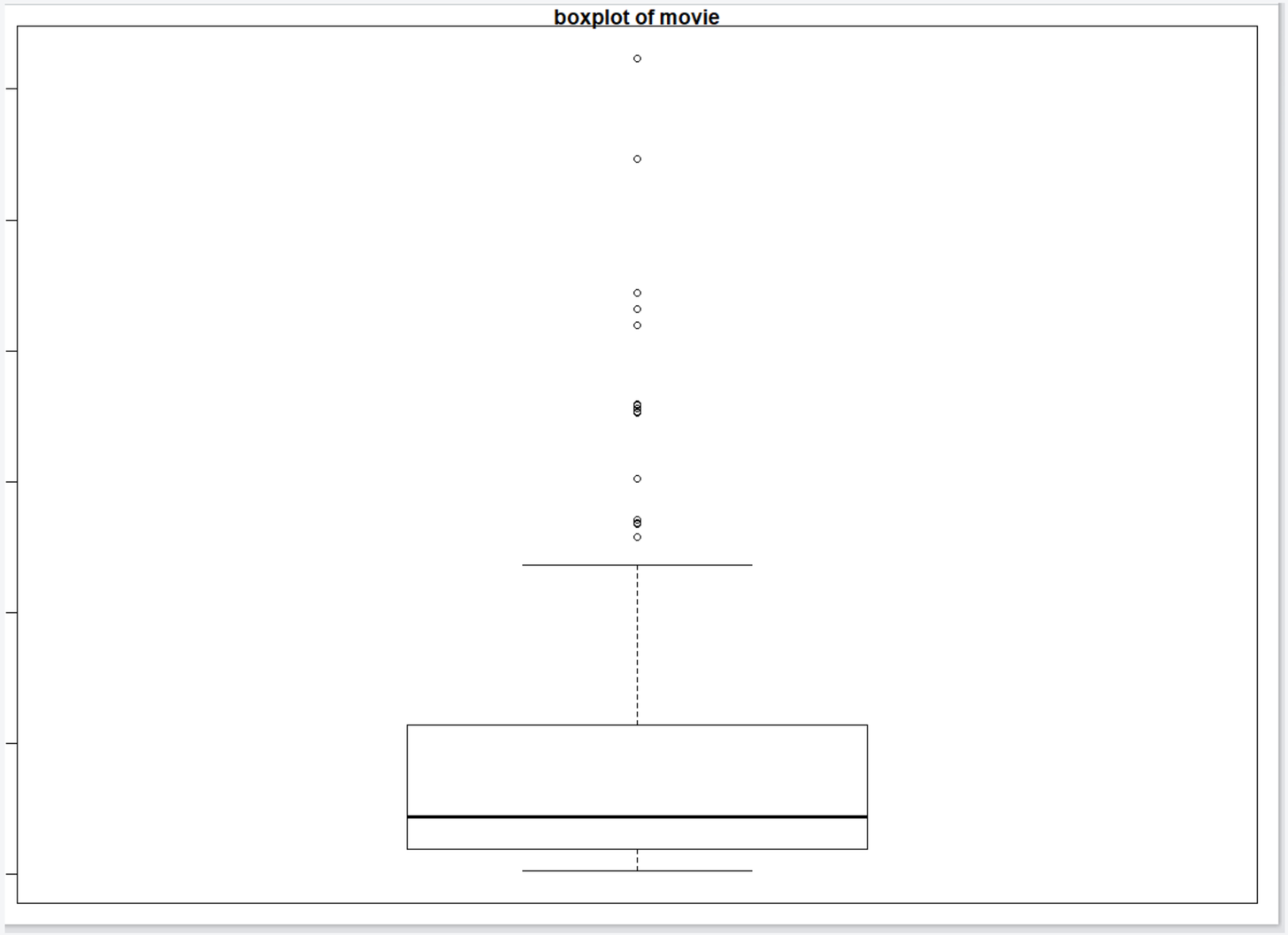
STEP 2.

Basic Statistical Analysis

Because We have too many independent variables and we do not know which variables are affecting the cumulative number of movie viewers, we should choose an independent variable with reference to the correlation coefficient.

It is good to choose the independent variables which has the correlation coefficient from 0.3 to 0.7.
(reference to blog)

We can get the correlation coefficient between the cumulative number of 1st week cinema audience and other variables by function ‘cor()’ in R.



Box plot

This is a boxplot about the cumulative number of 1st week cinema audience.

There are many outliers, so it help us to process the data.

Because there are many outliers, if K It was confirmed that the small outline would be sensitive to the outliers and the accuracy could be lowered.

Choose independent value

We concluded the it is right to analyze 6 variables except the variables which is too small absolute values like country, limitation of age, genre

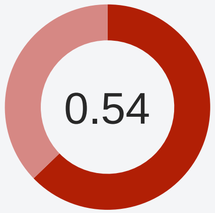
TV	Korea	US	Jepan	X12up	X15up	X19up	distributo	action	comedy	drama
-0.04452	0.16322	-0.09087	-0.09434	0.07917	0.06734	-0.14137	0.20273	0.2118	-0.02854	-0.07665

hiswar	crime	X1st_gold	X2nd_gol	naver_goc	naver_bad	naver_pe	blog	actor	director
0.01486	0.14864	0.09227	0.14862	0.6942	0.54174	0.50263	0.45136	0.47028	0.33235

Six meaningful independent variables

0.3~0.7

naver_good



naver_bad

naver_pe

blog

actor

director

STEP 3.

Knn

We classify the cumulative number of 1st cinema audience by the class , like High, Medium, Low, to implement the machine learning , Knn and decision tree.
The standard of classifying the class is calculated referring to the cumulative number of 1st cinema audience in the Internet news articles and five-number summary of it.

NEW PROJECT

	A	B	C	D	E	F	G	H	I	J	K
1	X1st_pe	Country	Agelimit	distributor	genre	naver_goo	naver_bad	naver_pe	blog	actor	director
2	M	0	2	0	1	6919	866	2339.34	199	0	0
3	M	2	2	0	4	3751	383	1470.5	158	0.0238	0
4	M	2	1	1	1	7337	677	5093.34	245	0.0938	0
5	L	1	1	1	3	12208	1071	3652.4	382	0.1333	0
6	L	1	3	1	8	4570	2003	4794.87	227	0.0588	0.067
7	L	2	1	0	1	2439	350	1038.86	76	0	0
8	M	1	3	0	1	21361	4102	12298.51	380	0.08	0.333
9	H	1	2	1	2	46891	2492	6706.54	230	0.3333	0
10	M	1	2	1	3	15918	2041	4407.66	156	0.3227	0
11	M	2	2	0	3	5068	759	8507.7	60	0	0
12	L	1	2	1	3	8640	1340	1991.06	179	0.1103	0
13	L	1	2	0	3	6385	658	4189.34	136	0.0238	0
14	L	1	0	1	8	3583	625	3952.55	281	0.0526	0.278
15	L	2	1	1	1	2374	328	2752.68	83	0.1364	0
16	H	1	2	1	3	15570	2644	7205.28	470	0.2788	0.25
17	L	2	2	0	1	4457	1042	7299.88	237	0.0862	0

2nd_pre-processing

We pre-processed the data again to perform the knn analysis.

```r
movie<-read.csv('movie.csv')
str(movie)
head(movie)
movie$X1st_pe<-factor(movie$X1st_pe,levels = c("H","M","L"))
summary(movie$X1st_pe)
summary(movie[,-1])

normalize<-function(x){return((x-min(x))/(max(x)-min(x)))}
movie_n<-as.data.frame(lapply(movie[2:11],normalize))
summary(movie_n)

set.seed(999)
train_sample<-sample(328,228)
movie<-movie[-5]
movie<-movie[-4]
movie<-movie[-3]
movie<-movie[-2]
movie_train1<-movie[train_sample,-1]

movie_test1<-movie[-train_sample,-1]

movie_train_labels1<-movie[train_sample,1]

movie_test_labels1<-movie[-train_sample,1]


library(class)

movie_test_pred<-knn(train = movie_train1, test = movie_test1,cl=movie_train_labels1,k=114)

library(gmodels)
CrossTable(x=movie_test_labels1,y=movie_test_pred,prop.chisq = F)
confusionMatrix(movie_test_labels1,movie_test_pred, positive = "L")
```

Knn R-code

In order to use Knn about this data, We set the class 'H' by more than two million, the class 'M' from 700 thousand, and the class 'L' by less than 700 thousand.

Result about using Knn

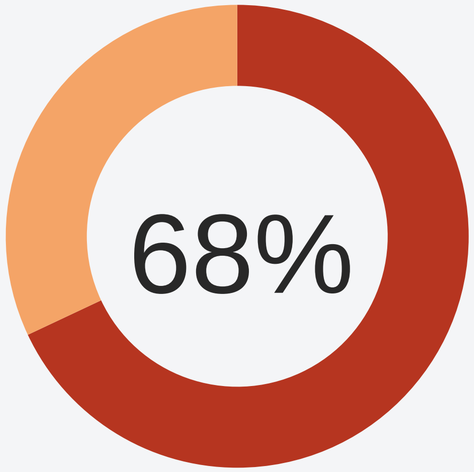
Cross table →
↓ Accuracy

```
          Reference
Prediction  H  M  L
         H  0  4  0
         M  0  6 14
         L  0  0 76

Overall Statistics

               Accuracy : 0.82
                 95% CI : (0.7305, 0.8897)
```

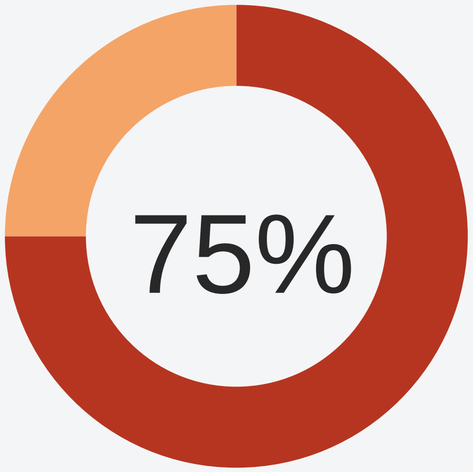
Total Observations in Table: 100

movie_test_labels1	movie_test_pred M	L	Row Total
H	4	0	4
	1.000	0.000	0.040
	0.400	0.000	
	0.040	0.000	
M	6	14	20
	0.300	0.700	0.200
	0.600	0.156	
	0.060	0.140	
L	0	76	76
	0.000	1.000	0.760
	0.000	0.844	
	0.000	0.760	
Column Total	10	90	100
	0.100	0.900	

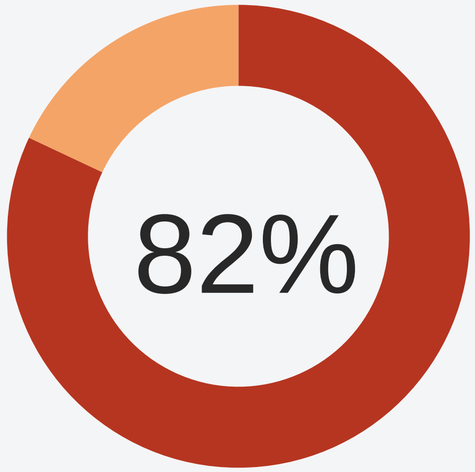
Accuracy depending on choosing different K.



K =35



K = 70



K = 114

NEW PROJECT

STEP 4.
Decision Tree

Decision tree R-Code

```
movie <- read.csv("movie.csv", stringsAsFactors = FALSE)

str(movie)

set.seed(999)

train_sample <-sample(328,228)
str(train_sample)

movie<-movie[-5]
movie<-movie[-4]
movie<-movie[-3]
movie<-movie[-2]

movie_train <- movie[train_sample,-1]
movie_test <- movie[-train_sample, ]

movie_train_label<- movie[train_sample, 1]
movie_train_label<- factor(movie_train_label)
summary(movie_train_label)
#install.packages("C50")
library(C50)

movie_model <- C5.0(movie_train,movie_train_label)

summary(movie_model)

movie_pred<-predict(movie_model,movie_test)

library(gmodels)
CrossTable(movie_test$X1st_pe,movie_pred,prop.chisq=FALSE,prop.c = FALSE,prop.r = FALSE,dnn= c('actual default','predicted default'))

install.packages('caret')
library(caret)
confusionMatrix(movie_test$X1st_pe,movie_pred, positive = "L")
```

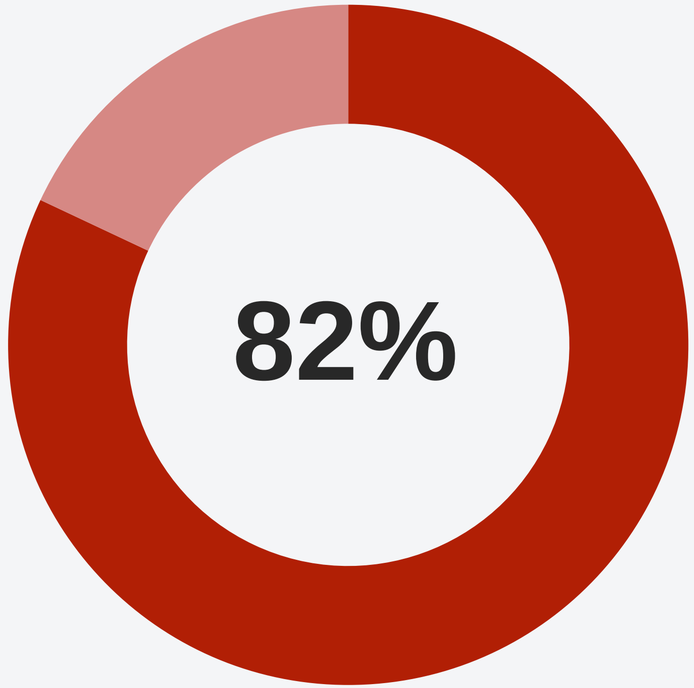
Result using decision tree

Cross table

actual default	predicted default H	M	L	Row Total
H	3 0.030	1 0.010	0 0.000	4
M	2 0.020	13 0.130	5 0.050	20
L	0 0.000	10 0.100	66 0.660	76
Column Total	5	24	71	100

Accuracy of decision tree

```
Confusion Matrix and Statistics

          Reference
Prediction  H  L  M
         H  3  0  1
         L  0 66 10
         M  2  5 13

Overall Statistics

               Accuracy : 0.82
                 95% CI : (0.7305, 0.8897)
```



Comparison of Knn and DecisionTree

	KNN	DESION TREE
H(4)	0/4	3/4
M(20)	6/20	13/20
L(76)	76/76	66/76

knn selects neightbor based on Euclidean distance.
By the way Since there are many outliers, we could not properly classify the H data included in outliers because we got high K

On the other hand, since the decision tree is not so, the data included in H is classified well.
However, if the tree is too large and detailed, the accuracy is lowered and the accuracy of M and L is somewhat lower than that of Knn.

CONCLUSION

We can predict the cumulative number of 1st cinema audience by 6 independent variables. However, it is not exact value but only the class. So, by regression analysis We can get the regression formula and analyze the exact value.

Step 4.

```r
#데이터셋 로드
moviedata <- read.csv(file.choose(), header = T, stringsAsFactors = T)
str(moviedata)

sum(is.na(moviedata)) # missing value 확인

cor(moviedata) # 상관계수 확인, 상관계수가 너무 높으면 다중 공선성 문제 가능성이 있으므로
               # 상관관계가 높은 변수는 독립변수로 사용불가
install.packages("car")
library(car)

fit <- lm(formula = X1st_pe ~ naver_good+naver_bad+naver_pe+blog+actor+director, data=moviedata)
vif(fit) # vif 가 10이 넘는것이 없음, 즉 다중공선성에 대한 문제가없음
summary(fit) # R-squared가 0.5847, 1주차 누적관객수에 대한 독립변수들의 설명력을 의미함
# p-value가 0.005이하이면 통계적으로 유의미한 회귀방정식인데, 지금 이식의 p-value는 0.00000016으로 유의미
```

Regression Analysis

Regression Code

```
> vif(fit) # vif 가 10이 넘는것이 없음, 즉 다중공선성에 대한 문제가없음
naver_good  naver_bad   naver_pe       blog      actor   director
  2.517230   2.326679   1.905457   1.334089   1.138624   1.211298
> summary(fit) # R-squared가 0.5847,  1주차 누적관객수에 대한 독립변수들의 설명력을 의미함

Call:
lm(formula = X1st_pe ~ naver_good + naver_bad + naver_pe + blog +
    actor + director, data = moviedata)

Residuals:
     Min       1Q   Median       3Q      Max
-1603681  -325599   -57044   209394  2421173

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.325e+04  6.916e+04  -0.481  0.63106
naver_good   4.446e+01  5.630e+00   7.897 8.42e-14 ***
naver_bad   -1.358e+01  4.288e+01  -0.317  0.75178
naver_pe     1.477e+01  9.415e+00   1.569  0.11782
blog         1.007e+03  3.426e+02   2.940  0.00358 **
actor        3.078e+06  4.966e+05   6.198 2.27e-09 ***
director     8.516e+05  4.123e+05   2.066  0.03988 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 617900 on 256 degrees of freedom
Multiple R-squared:  0.5942,    Adjusted R-squared:  0.5847
F-statistic: 62.47 on 6 and 256 DF,  p-value: < 2.2e-16

>
```

Result using regression analysis

Looking at the first line of the output screen, we checked the Vif value to determine if there is a multicollinearity problem.

Multi-collinearity refers to a problem that affects regression analysis negatively because of high correlation between independent variables.

The next Vif value is called the Dispersion Expansion Factor, which is a value that determines whether there is a high correlation between independent variables and can range from 1 to infinity.

If this Vif value exceeds 10, it is judged that there is a problem in multi-collinearity.

Regression formula

$$Y_{(\text{1주차 누적 관객수})} = 44.46\text{naver\_good} - 13.58\text{naver\_bad} + 14.77\text{naver\_pe} + 1007\text{Blog} + 30780100\text{Actor} + 851600\text{Director} - 33250$$

R – squard ?

Definition

"R-Suared" is the square of the correlation coefficient between variables. This value is more than 0.6 in academia and 0.4 in marketing research It is interpreted as meaningful.

Meaning

"R -square" is 0.58, which means that each independent variable accounts for 58.47% of the dependent variable. At first glance, this regression equation may seem insignificant because the explanatory power of the independent variable is less than 60%

Explain

Given that it is difficult to explain 10% of the social phenomena, 58.47% is not small. When regression analysis is performed, if the R squared value is less than 0.4, the remaining indicator is not necessary to see and meaningless.

	A	B	C	D	E	F	G	H	I
1	실제값	예측값	실제값/예	70%예측	naver_bad	naver_pe	blog	actor	directo
2	1,344,283	522,507	2.572756	No	-9.676	13.87	1017	3000000	826
3	1,196,692	406,993	2.940326	No	-9.676	13.87			
4	2,009,291	908,903	2.210677	No	-9.676	13.87			
5	415,520	1,354,871	0.306686	Yes	-9.676	13.87			
6	694,266	703,597	0.986738	No	-9.676	13.87			
7	610,123	189,446	3.220564	No	-9.676	13.87			
8	1,192,642	1,956,935	0.609444	Yes	-9.676	13.87			
9	5,469,383	3,337,956	1.638543	No	-9.676	13.87			
10	2,183,079	1,855,741	1.176392	No	-9.676	13.87			
11	1,189,551	387,197	3.072211	No	-9.676	13.87			
12	320,108	898,531	0.356257	Yes	-9.676	13.87			
13	287,965	534,274	0.538984	Yes	-9.676	13.87			
14	373,983	872,970	0.428403	Yes	-9.676	13.87			
15	335,848	626,921	0.53571	Yes	-9.676	13.87			
16	3,024,092	2,267,781	1.333503	No	-9.676	13.87			
17	632,425	779,731	0.811081	No	-9.676	13.87			
18	1,428,021	1,441,843	0.990414	No	-9.676	13.87			
19	237,196	586,164	0.404658	Yes	-9.676	13.87	1017	3000000	826
20	546,764	1,359,052	0.402313	Yes	-9.676	13.87	1017	3000000	826
21	1,977,222	1,746,287	1.132243	No	-9.676	13.87	1017	3000000	826

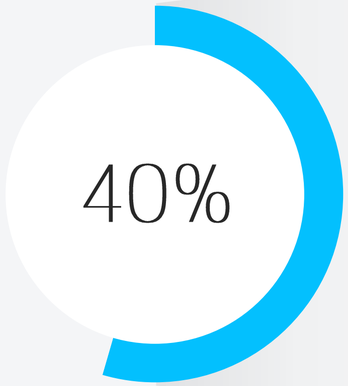
Analysis result

We express the predicted value by regression formula and accurate value., If value obtained by dividing the actual value into the forecast value is greater than 0.7 or less than 1.3, it is 'Yes'. Otherwise, it is 'No'.

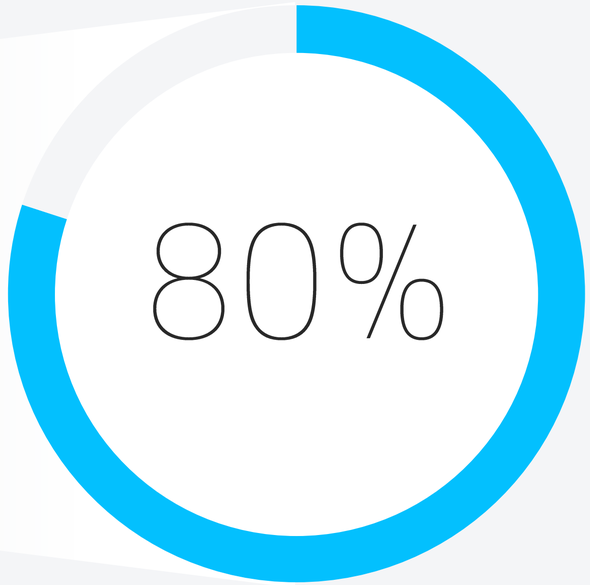
Then we checked the accuracy.

problem

Accuracy is very low.
Because We think we need to a standard We choose 'Yes' and analyze it.
If the variables about actors and directors is near to 0 or the variables about the number of blog postings is near to 0, then predicted value is not exact.
So we need to choose predicted model which is appropriate to the value about actors, director, and blog postings.



The percentage of 'Yes'
The reason why the percentage of 'Yes' is low is because of actors, directors, and blog postings.

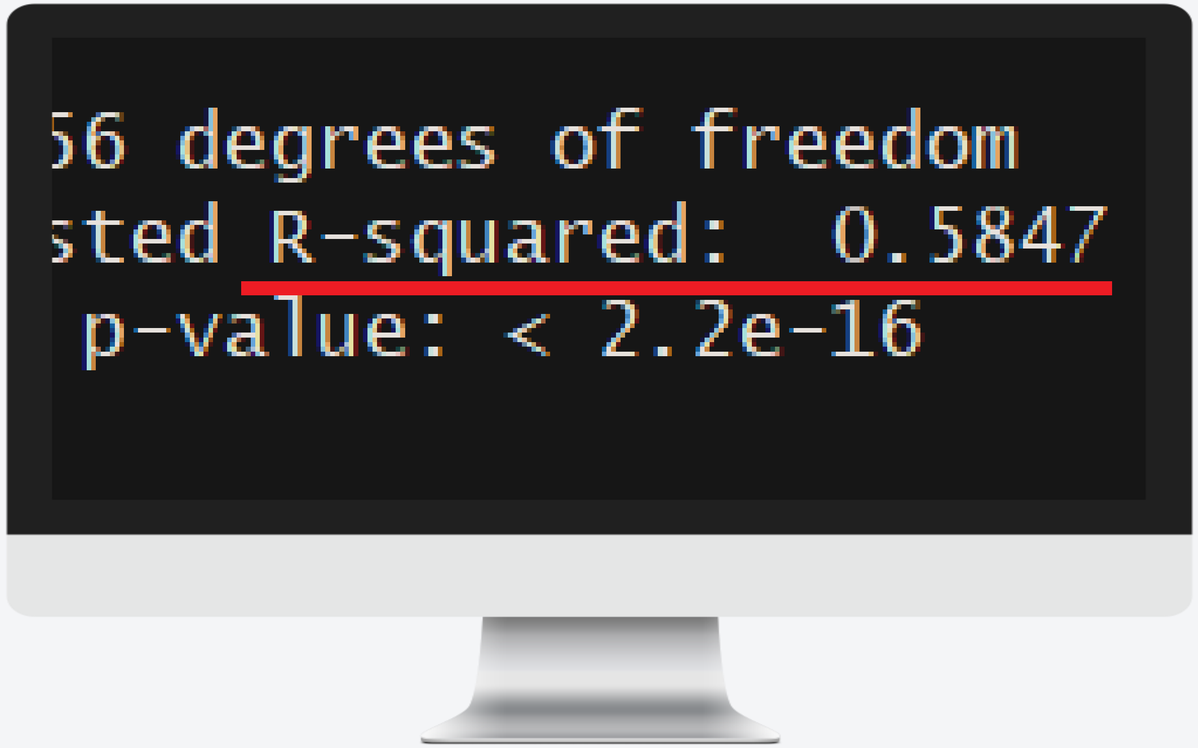


If the model has the variables about actors and directors which is not near to 0 or the variables about blog postings which is greater than 50, then it is a good predictive model.
Since Most of Korean films are satisfied with this standard, there is no big problem to analyze it by this formula.

Step 6.

apply

R-squard



We choose some films that more than 50 posting in blog are uploaded (one week before the opening week) and influence of actors and directors is not 0. We judged that the model is accurate.
Since R-squared is 0.58, it can be interpreted that it is quite reasonable to predict social phenomenon.

example

Crime Town

Release 2017.10.03

diretor : 강윤성 actor : 윤계상 마동석

Taxi Driver

Release 2017.08.02

diretor : 장훈 actor : 송강호

The Swindlers

Release : 2017.11.22

diretor :장창원 actor : 현빈 , 유지태

example

I can speak

Release : 2017.09.21

diretor : 김현석, actor : 나문희 이제훈

Let me eat your pancreas

Release :2017.10.25

diretor : 츠키카와 쇼 actor : 하마베,키타무라

If only

Release : 2017.11.29

diretor :길 정거 ,

actor :제니퍼 러브 휴잇, 폴 니콜스

예측값 정리(오름차순)		
영화제목	실제값	예측값
너의 췌장을 먹고싶어	21,004	336,797
이프온리	179,177	923,099
아이캔스피크	1,099,933	2,491,129
꾼	2,199,937	3,195,798
범죄도시	2,388,597	2,427,659
택시운전사	5,812,815	5,839,862

http://www.kobis.or.kr/kobis/business/mast/mvie/searchMovieList.do

The values for each independent variable were treated as a result of the discoveries found in Naver blogs and sites.
There were quite a few similar results, and there were many different results.
The reason was highly influenced since the coefficient about actors and directors is too high in regression formula.
The film's actors and directors did not have enough film works to deduct accurate predictions.

YES

Is the director's influence appropriate?

YES

Is the actor's influence appropriate?

NO

New actor or new director without work

NO

Lack of blog postings

I felt that…

I learned bigdata processing through only paper. But this time, I learned practical skills about it inR directly. I already learned R programming, but I only learned the basic skills of R. Then I didn't know how to apply R to the real work and it wasn't touched to me. In that sense, this lecture is very beneficial to learn various analytical methods

by _ park ji su

During this semester, while taking a big data processing class. It was nice to learn various classification techniques and analysis techniques. I think it is a good opportunity to have a good base for R language..

by – Lee u seok

I felt that…

Before I took this lecture, I already studied the basic skills in R. Although I just heard that R is very useful, i didn't know how to apply R in the real world. However, through this lecture I knew how to use R in real work and studied the various classification methods. I thought that classification methods are very difficult. But there are many simple machine learning like Knn and skyline. It was surprising. Through the project, it's good to apply the machine learning to some data.

by – hwang se young

Last year, when i was learning R, I just wrote code. I did not know what it meant. However, during this semester, this class was very informative lecture that learned various analysis methods and its meaning.

by – jang hee ju

Question

thank you