Cluster Analysis
보고서

2011100038 박병진
2011100074 김혁준
2012100005 강나루

Contents

Titanic

TITANIC

√ 미국 뉴욕을 향하던 4월 15일 새벽, 북대서양의 빙하를 만나 침몰

√ 최초 출항 시 탑승했던 2224명의 승선자 중 1514명이 사망

침몰 당시 선박에는 어떤 사람들이 타고 있었을까?

또, 생존자들을 묶어주는 어떤 특징이 있을까?

Description

Description

변수명	설명	비고
Pclass	객실등급	1~3등석으로 표기
Name	승객이름	First name, Last name
Sex	성별	male/female로 표기
Age	나이	1세 미만 소수점표기
Sibsp	동승 형제자매	-
Parch	동승 부모,자녀	
Ticket	티켓번호	-
Fare	가격	화폐단위 : british pound
Cabin	갑판	A~G갑판으로 배정
Embarked	승선지	C=Cherbourg, France Q = Queenstown, Ireland S = Southampton, England
Boat	구명보트	구조자 중 탑승보트번호
Body	식별번호	사망자의 식별번호로 추정
Home.dest	고향/목적지	-

Scaling

① 데이터 선별

√ 주어진 데이터셋은 **다양한 정보**를 포함

– 하지만 데이터 표현 형식 혹은 결측치로 인한 표본의 불균형 등으로 인해 사용할 데이터 선별 작업 요구.

variable	cause
Name	분석에 필요한 자료라고 생각되지 않아서 삭제. 분석 후 참고용으로 사용.
Ticket	티켓번호도 분석에 중요하지 않은 변수라 판단.
Cabin	관측치가 295개로 전체 데이터 1309개에 비해서 부족하여 분석에 부적합하다고 판단 but, 객실구조에 의한 생존율 비교 필요.
Boat	데이터 입력 정보형태가 분석하기 어려웠고 데이터의 개수도 486개로 결측치 많이 존재.
Body	관측치가 121개라서 분석이 어렵고, 분석에 필요하지 않은 정보라 판단.
Home.dest	데이터 분석이 어렵고 장소도 다양해서 분석이 어려움.

Scaling

② 가변수 생성

√ 자료에 범주형 자료 존재

–〉 가변수(dummy variable) 생성 필요.

– sex의 경우 male/female, embark의 경우 S(Southampton), C(Cherbourg), Q(Queenstown)로 범주화 되어 있음.

```
> library(dummies)
> embarked.dum=dummy(titanic$embarked)
> titanic=cbind(titanic,embarked.dum)
> sex.dum=dummy(titanic$sex)
> titanic=cbind(titanic,sex.dum)
```

```
> head(titanic)
  pclass                                            name    sex age sibsp parch   ticket     fare
1      1                          Artagaveytia, Mr. Ramon   male  71     0     0 PC 17609  49.5042
2      1                           Astor, Col. John Jacob   male  47     1     0 PC 17757 227.5250
3      1 Astor, Mrs. John Jacob (Madeleine Talmadge Force) female  18     1     0 PC 17757 227.5250
4      1                       Aubart, Mme. Leontine Pauline female  24     0     0 PC 17477  69.3000
5      1                            Baxter, Mr. Quigg Edmond   male  24     0     1 PC 17558 247.5208
6      1   Baxter, Mrs. James (Helene DeLaudeniere Chaput) female  50     0     1 PC 17558 247.5208
    cabin embarked boat body          home.dest embarkedC embarkedQ embarkedS sexfemale sexmale
1                C        22 Montevideo, Uruguay         1         0         0         0       1
2 C62 C64        C       124        New York, NY         1         0         0         0       1
3 C62 C64        C    4  155        New York, NY         1         0         0         1       0
4     B35        C    9  155       Paris, France         1         0         0         1       0
5 B58 B60        C       155        Montreal, PQ         1         0         0         0       1
6 B58 B60        C    6  155        Montreal, PQ         1         0         0         1       0
>
```

Scaling

③ 결측값 대체

√ 데이터셋에 결측치 多

age

– 결측치 263개
– 연속형 자료이므로 대표적 imputation 기법인 median으로 대체.

fare

– 결측치가 1개이지만 영향이 큰 자료일 수도 있어서 자료검색. =〉 3등급 객실이라 median 대체해도 자료에 큰 영향 없을 것이라 판단.

embark

– 결측치 2개.
– 출발지가 S(Southampton)이므로 S일 가능성이 가장 높다고 생각.(S가 mode)

Scaling

④ 변수 표준화

√ 변수들마다 척도(scale)가 다양

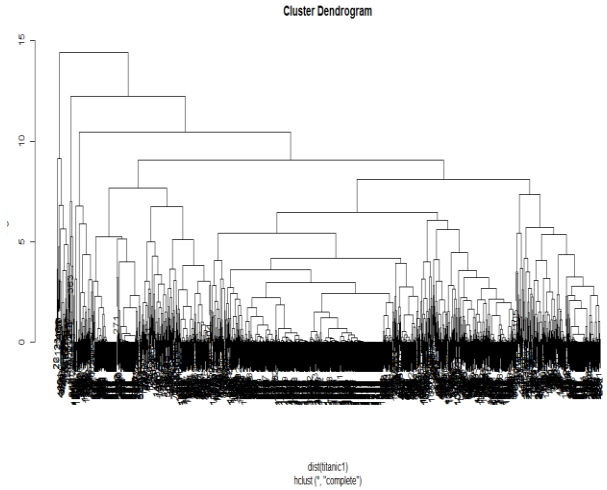
– 분석을 시작하기 전에 변수들간의 척도를 표준화 시켜줄 필요가 있음.

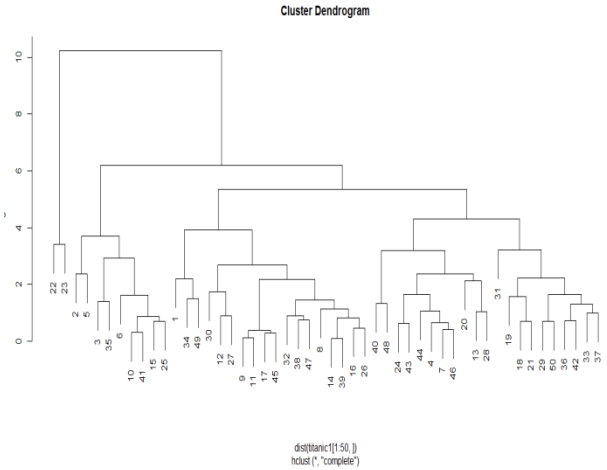
```
> titanic1=as.data.frame(scale(titanic1))
```

```
> head(titanic1)
     pclass        age      sibsp      parch      fare embarkedC  embarkedQ embarkedS  sexfemale
1 -1.545507  3.2155010 -0.4789037 -0.4448295 0.3135416   1.96092 -0.3219173 -1.526109 -0.7432129
2 -1.545507  1.3557914  0.4811039 -0.4448295 3.7541222   1.96092 -0.3219173 -1.526109 -0.7432129
3 -1.545507 -0.8913577  0.4811039 -0.4448295 3.7541222   1.96092 -0.3219173 -1.526109  1.3444816
4 -1.545507 -0.4264303 -0.4789037 -0.4448295 0.6961320   1.96092 -0.3219173 -1.526109  1.3444816
5 -1.545507 -0.4264303 -0.4789037  0.7104916 4.1405780   1.96092 -0.3219173 -1.526109 -0.7432129
6 -1.545507  1.5882551 -0.4789037  0.7104916 4.1405780   1.96092 -0.3219173 -1.526109  1.3444816
     sexmale
1  0.7432129
2  0.7432129
3 -1.3444816
4 -1.3444816
5  0.7432129
6 -1.3444816
>
```

Hierarchical clustering

군집의 개수를 정하기 위해 계층분석 시행

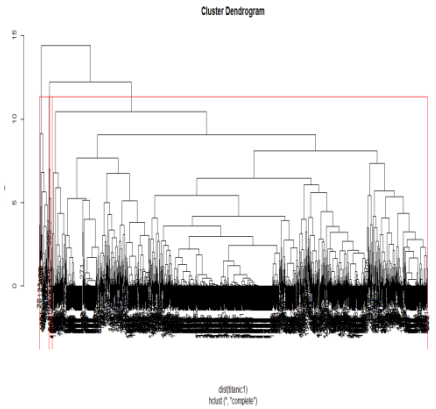


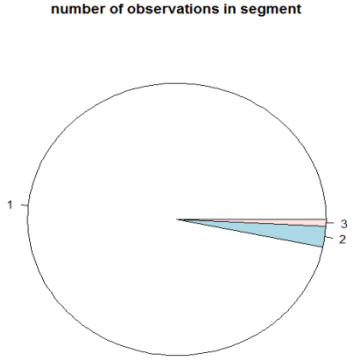



```
> h.mean
      pclass         age        sibsp       parch       fare    embarkedC    embarkedQ   embarkedS
1  0.03175252 -0.02676962 -0.009516113 -0.07341112 -0.1012082  0.005599333  0.008227891 -0.01018059
2 -1.54550719  0.84001257  0.271102230  0.74659533  3.9503906 -0.046357175 -0.321917258  0.24584764
3  0.84159477  0.63726724  0.306557057  6.27703839  0.1561004 -0.509575051 -0.010468854  0.45649981
    sexfemale     sexmale
1 -0.01763184  0.01763184
2  0.56159620 -0.56159620
3  0.39552959 -0.39552959
> dist(h.mean,method="euclidean",diag=TRUE)
         1        2        3
1 0.000000
2 4.610198 0.000000
3 6.512761 7.150815 0.000000
```

Hierarchical clustering

군집의 개수를 정하기 위해 계층분석 시행

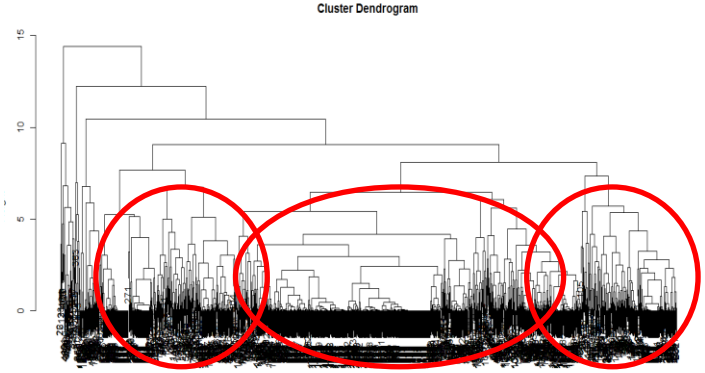




√ 군집1에 비해 2,3의 **데이터량이 부족**한 점 파악

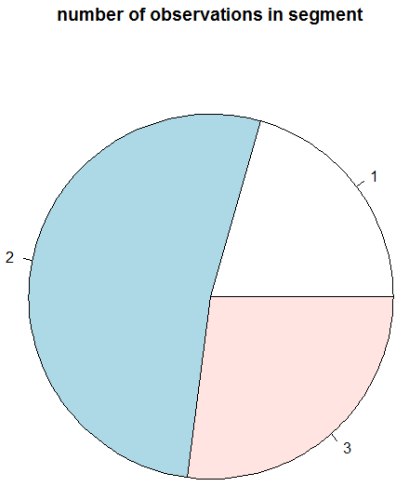
– 계층군집분석의 특징때문에 처음에 묶인 군집에서 이동이 불가해서 생긴 문제로 파악

=〉 Dendrogram으로 군집수가 3개인 것만 파악함.



K-means clustering

K=3으로 k-means 군집분석 시행

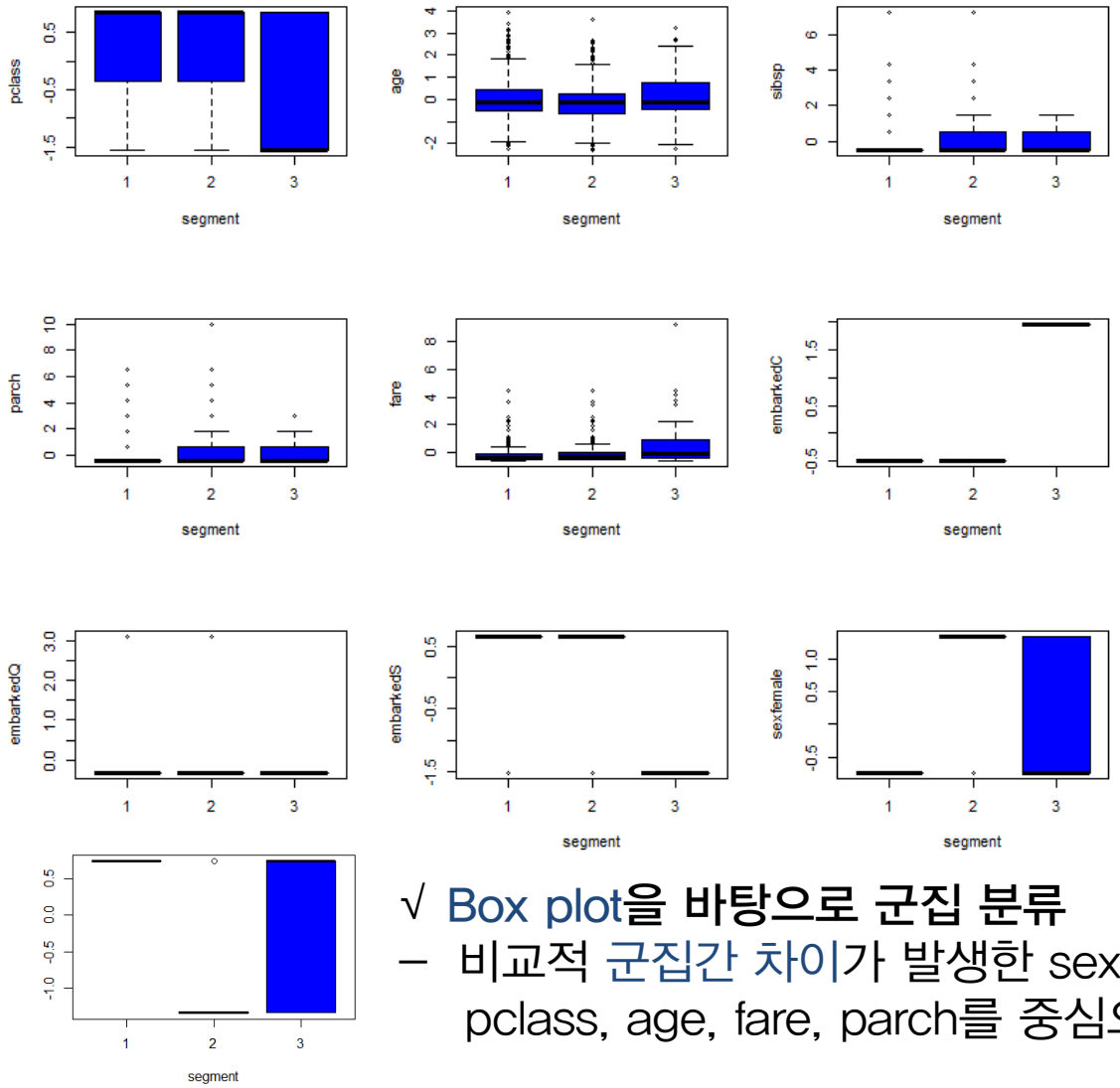



```
> dist(k.clust$centers, method = "euclidean", diag = TRUE)
         1        2        3
1 0.000000
2 3.586223 0.000000
3 3.677905 3.039933 0.000000
```

```
> k.clust$centers
      pclass         age       sibsp       parch        fare  embarkedC  embarkedQ  embarkedS
1 -0.52877858  0.147100700 -0.09490068 -0.01693281  0.56154502  1.9609203 -0.3219173 -1.5261088
2  0.19341818  0.007802844 -0.06266684 -0.17834668 -0.23964535 -0.5095751 -0.0068315  0.4541844
3  0.02903605 -0.127294172  0.19364399  0.35802072  0.03542347 -0.5095751  0.2587493  0.2851218
   sexfemale    sexmale
1  0.1305259 -0.1305259
2 -0.7432129  0.7432129
3  1.3385842 -1.3385842
```

K-means clustering

Box plot

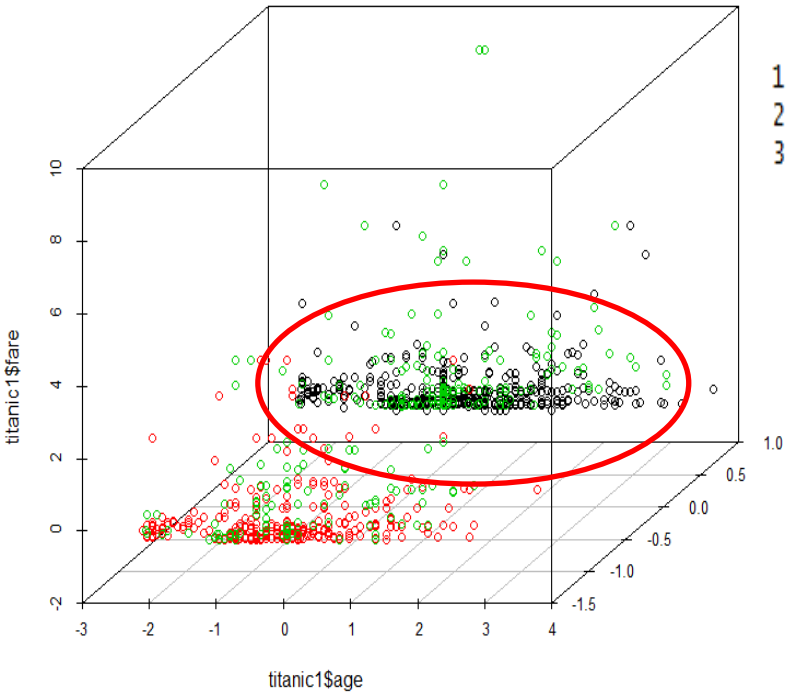


√ Box plot을 바탕으로 군집 분류

– 비교적 군집간 차이가 발생한 sex, embark, pclass, age, fare, parch를 중심으로 분석.

K-means clustering

Cluster1 : 홀로 탑승한 남자승객



	pclass	age	sibsp	parch
1	-0.52877858	0.147100700	-0.09490068	-0.01693281
2	0.19341818	0.007802844	-0.06266684	-0.17834668
3	0.02903605	-0.127294172	0.19364399	0.35802072

	fare	embarkedC	embarkedQ	embarkedS
	0.56154502	1.9609203	-0.3219173	-1.5261088
	-0.23964535	-0.5095751	-0.0068315	0.4541844
	0.03542347	-0.5095751	0.2587493	0.2851218

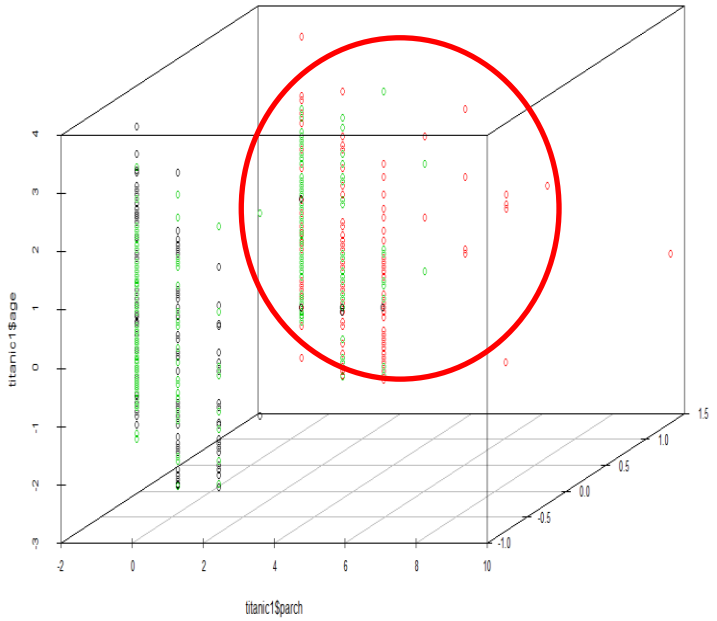
titanic1$sexmale

	sextemale	sexmale
1	0.1305259	-0.1305259
2	-0.7432129	0.7432129
3	1.3385842	-1.3385842

√ **Cluster 1**은 **요금**은 비교적 **적게** 내었으며, 대부분 **남성**으로 이루어지고 **가족이 없는** 돈을 벌기 위해 온 남성으로 파악됨.

K-means clustering

Cluster2 : 가족을 동반한 어린 여자승객



	pclass	age	sibsp	parch
1	-0.52877858	0.147100700	-0.09490068	-0.01693281
2	0.19341818	0.007802844	-0.06266684	-0.17834668
3	0.02903605	-0.127294172	0.19364399	0.35802072

	fare	embarkedC	embarkedQ	embarkedS
	0.56154502	1.9609203	-0.3219173	-1.5261088
	-0.23964535	-0.5095751	-0.0068315	0.4541844
	0.03542347	-0.5095751	0.2587493	0.2851218

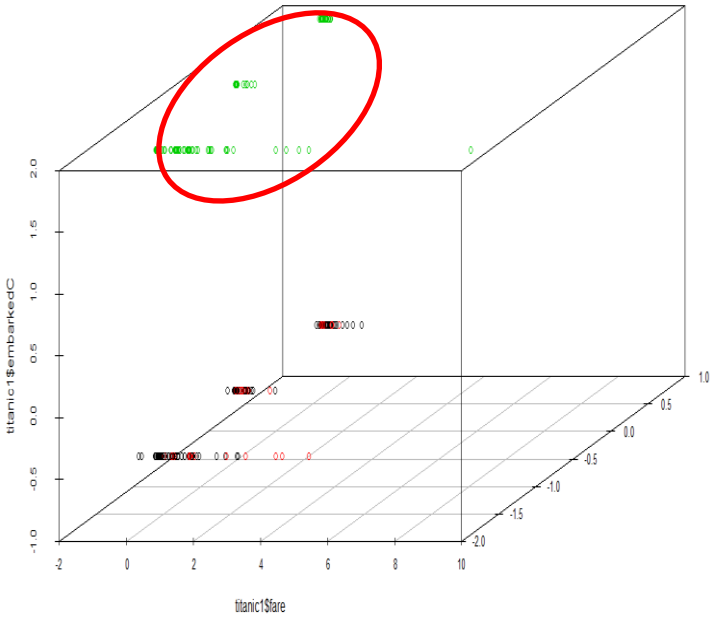
titanic1$sexfemale

	sexfemale	sexmale
1	0.1305259	-0.1305259
2	-0.7432129	0.7432129
3	1.3385842	-1.3385842

√ Cluster 2는 나이가 비교적 어리며, 대부분 여성으로 이루어지고 형제자매나 부양자식 혹은 부모님을 동반한 젊은 여성으로 파악됨.

K-means clustering

Cluster3 : 부유한 승객이 많은 프랑스 사람



	pclass	age	sibsp	parch
1	-0.52877858	0.147100700	-0.09490068	-0.01693281
2	0.19341818	0.007802844	-0.06266684	-0.17834668
3	0.02903605	-0.127294172	0.19364399	0.35802072

	fare	embarkedC	embarkedQ	embarkedS
	0.56154502	1.9609203	-0.3219173	-1.5261088
	-0.23964535	-0.5095751	-0.0068315	0.4541844
	0.03542347	-0.5095751	0.2587493	0.2851218

titanic1$pclass

	sextemale	sexmale
1	0.1305259	-0.1305259
2	-0.7432129	0.7432129
3	1.3385842	-1.3385842

√ Cluster 3는 요금을 많이 내었으며, 객실등급도 1~3등급으로 다양하게 이루어지고 프랑스에서 승선한 승객으로 파악.

Conclusion

인물비교

√ **Andersson, Mr. Johan Samuel** (26세/남성/3등석)

– 일용직 노동자, 영국인. 뉴욕에 거주하는 누나의 가족을 만나기 위해 타이타닉호에 탑승하였으나 사망. 시신발견되지 않음.

cluster1

√ **Mrs Emily Richards** (24세/여성/2등석)

– 사우스햄튼에서 두 아들과 남편과 함께 탑승.
친정어머니와 형제 자매들도 탑승
어머니와 자녀, 자매들과 함께 lifeboat 4를 타고 탈출함

cluster2

√ **Spedden, Mr. Frederic Oakley** (45세/남성/1등석)

– 미국인, 뉴요커, 하인을 거느린 부자.
가정부와 유모 동반해서 탑승.
3번 구명보트에 탑승해서 생존.

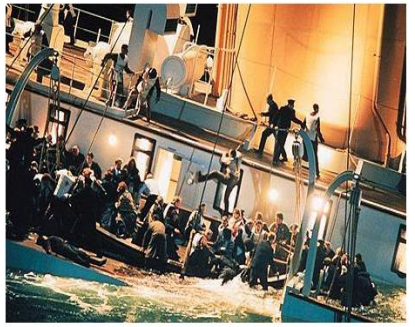
cluster3

Conclusion

생존율 비교

분류	탑승자	생존률	사망률	생존자	사망자
1등실, 어린이	6	83%	17%	5	1
2등실, 어린이	24	100%	0%	24	0
3등실, 어린이	79	34%	66%	27	52
1등실, 여성	144	97%	3%	140	4
2등실, 여성	93	86%	14%	80	13
3등실, 여성	165	46%	54%	76	89
1등실, 남성	175	33%	67%	57	118
2등실, 남성	168	8%	92%	14	154
3등실, 남성	462	16%	84%	75	387
승무원, 여성	23	87%	13%	20	3
승무원, 남성[44]	885	22%	78%	192	693
어린이 총합	109	51%	49%	56	53
여성 총합	425	74%	26%	316	109
남성 총합	1690	20%	80%	338	1352
전체 총합	**2224**	**32%**	**68%**	**710**	**1514**

"타이타닉 침몰시 1등석 생존율 더 높았다"



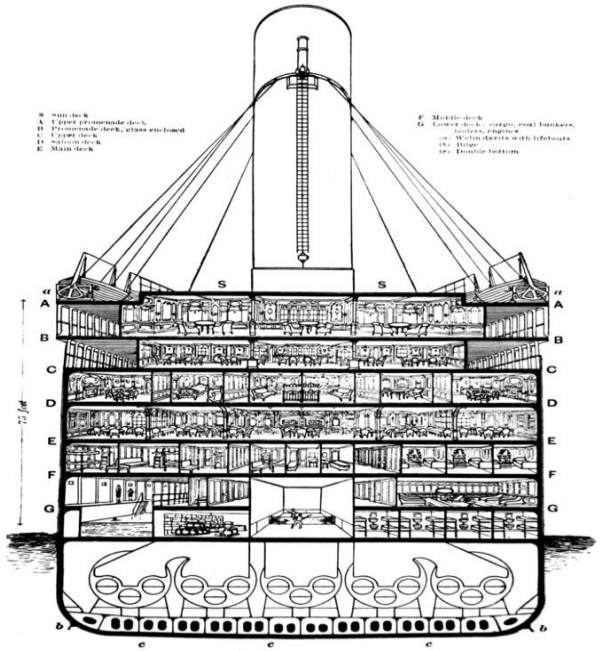
초호화 여객선 '타이타닉'이 침몰했을 당시, 1등석의 탑승객들이 3등석 탑승객들보다 더 많이 생존했던 것으로 밝혀졌다.

√ **1등실에 탔던 사람**과 **어린이와 여성**의 생존률이 다른 집단에 비해서 **높은** 것을 확인할 수 있음.

=>이를 통해서 유추해 보았을 때 어린이와 함께 동반한 젊은 여성이 속해있는 **군집2**와 비교적 부유하고 1등실이 많았던 **군집3**이 군집1에 비해 **생존율이 높았을 것**이라 예측가능.

Conclusion

생존율 비교



A갑판	1등실 객실, 라운지
B갑판	1등실 상당수
C갑판	1등실 객실, 2등실 도서관
D갑판	1,2,3등실 객실
E갑판	1,2,3등실 객실
F갑판	2등실 객실 일부, 3등실
G갑판	선원 등

√ cabin이 G갑판으로 갈수록 **아래**에 위치

- E갑판에서도 군집3에 속하는 인물들이 보이지만 F와 G갑판에서는 군집1과 군집2에 속하는 인물들로 구성되어 있음.
- Cabin의 관측치가 충분히 많았다면 이 역시 군집형성에 영향을 미쳤을 것이라 판단됨.