

# 8 장

---

## 타이타닉 호 데이터 분석

---

### 1 데이터 생성

---

3장의 1. import CSV - 5) Titanic Dataset를 참고해 데이터를 불러온다. 데이터베이스명은 mydata로 생성하고, 테이블은 dataset4로 설정하자.

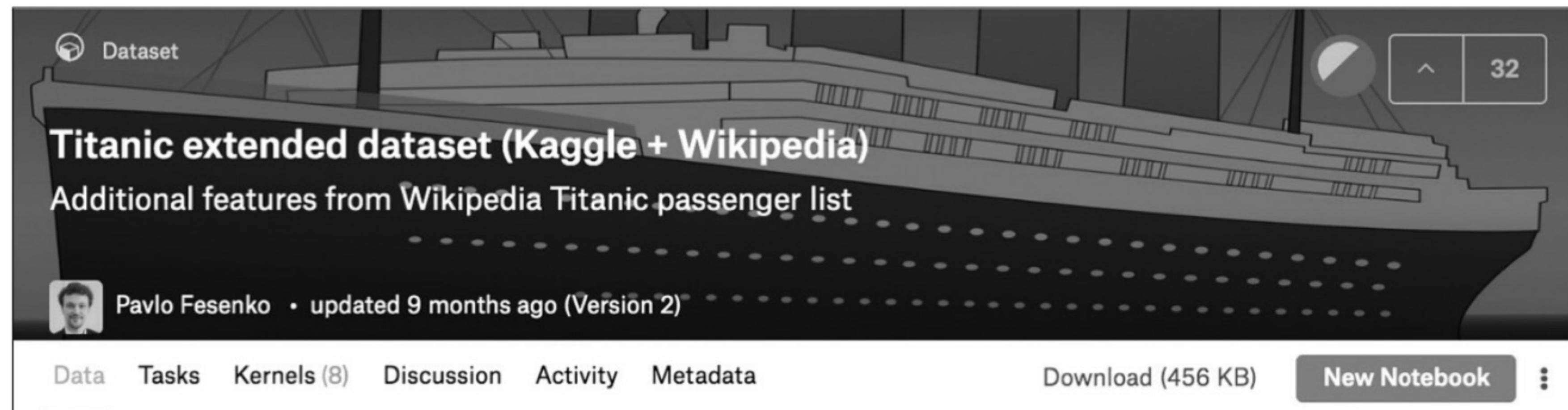
타이타닉 호 데이터 세트는 다음과 같이 구성되어 있다.

변수명	정의	키
survival	생존 여부	0 = No, 1 = Yes
pclass	티켓 클래스	1 = 1st, 2 = 2nd, 3 = 3rd
sex	성별	
age	사고 당시 연령	
sibsp	타이타닉호에 탑승한 형제/배우자 수	

변수명	정의	키
parch	타이타닉호에 탑승한 부모/자녀 수	
ticket	티켓 번호	
fare	Passenger Fare	
cabin	객실 번호	
embarked	승선 항구	C = Cherbourg, Q = Queenstown, S = Southampton
wikiid	위키 id	
name	이름	
age	연령	
hometown	고향	
boarded	출발지	
destination	목적지	
lifeboat	생존 보트	
body	body	
class	class	

데이터에 대한 상세 내용은 다음의 kaggle 주소에서 확인 가능하다.

<https://www.kaggle.com/pavlofesenko/titanic-extended>



[그림 8-1 Kaggle, Titanic Extended Dataset]

## 2 요인별 생존 여부 관계

타이타닉 데이터의 각 요인과 생존 여부에 어떤 관계가 있었는지 살펴보자.

### 1) 성별

먼저 타이타닉 데이터의 구조를 파악하기 위해 10개의 데이터만 출력해 구조를 살펴보자.

```
SELECT *
FROM MYDATA.DATASET4 LIMIT 10;
```

PASSENGERID	SURVIVED	PCLASS	NAME	SEX	AGE	PARCH	TICKET
1	1	0	3 Braund, Mr. Owen Harris	male	22	0	A/5 21171
2	2	1	1 Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	0	PC 17599
3	3	1	3 Heikkinen, Miss. Laina	female	26	0	STON/O2. 3101282
4	4	1	1 Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	0	113803
5	5	0	3 Allen, Mr. William Henry	male	35	0	373450
6	7	0	1 McCarthy, Mr. Timothy J	male	54	0	17463
7	8	0	3 Palsson, Master. Gosta Leonard	male	2	1	349909
8	9	1	3 Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	2	347742
9	10	1	2 Nasser, Mrs. Nicholas (Adele Achem)	female	14	0	237736
10	11	1	3 Sandstrom, Miss. Marguerite Rut	female	4	1	PP 9549

[그림 8-2 쿼리 실행 결과]

데이터를 보면, PassengerId라는 칼럼을 확인할 수 있다. PassengerId를 카운트하면, 승객 수를 집계할 수 있다. 이제 PassengerId에 중복이 존재하는지 확인해 보자.

```
SELECT COUNT(PASSENGERID) N_PASSENGERS,
COUNT(DISTINCT PASSENGERID) N_D_PASSENGERS
FROM MYDATA.DATASET4;
```

N_PASSENGERS	N_D_PASSENGERS
1	714

[그림 8-3 쿼리 실행 결과]

단순히 PassengerId를 카운트한 결과와 중복을 제거하고 카운트한 결과가 일치하는 것을 확인할 수 있다. 즉 PassengerId 값에는 중복이 존재하지 않는다. 이제 성별에 따른 승객 수와 생존자 수를 구해 보자.

```
SELECT SEX,  
COUNT(PASSENGERID) N_PASSENGERS,  
SUM(SURVIVED) N_SURVIVED  
FROM MYDATA.DATASET4  
GROUP  
BY 1  
;
```

SEX	N_PASSENGERS	N_SURVIVED
1 male	453	93
2 female	261	197

[그림 8-4 쿼리 실행 결과]

생존자 수를 계산할 때, SURVIVED의 합으로 생존자 수를 계산했다. SURVIVED(생존 여부)에서 1은 생존, 0은 사망을 의미하므로, SURVIVED의 합을 구하면 생존자 수를 계산할 수 있다.

결과를 살펴보면, 여성 생존자 수가 남성보다 높다. 이제 성별 탑승객 수와 생존자 수의 비중을 구해 보자.

```
SELECT SEX,  
COUNT(PASSENGERID) N_PASSENGERS,  
SUM(SURVIVED) N_SURVIVED,  
SUM(SURVIVED)/ COUNT(PASSENGERID) SURVIVED_RATIO  
FROM MYDATA.DATASET4  
GROUP  
BY 1  
;
```

SEX	N_PASSENGERS	N_SURVIVED	SURVIVED_RATIO
1 male	453	93	0.2052980
2 female	261	197	0.7547893

[그림 8-5 쿼리 실행 결과]

결과를 정리해 보면 남성이 여성보다 타이타닉호에 더 많이 탑승했고, 생존율은 여성이 남성보다 약 55%p 높았다.

## 2) 연령, 성별

앞에서 성별에 따른 탑승객 수, 생존자 수, 생존율을 계산해 보았다. 이제 연령에 따른 생존율을 살펴보자.

5장에서 살펴본 것처럼 연령은 10세 단위로 나누어 살펴보겠다. 10세 단위로 연령을 나누는 방법은 다음과 같다.

```
SELECT FLOOR(AGE/10)*10 AGEBAND,  
AGE  
FROM MYDATA.DATASET4
```

AGEBAND	AGE
1	20
2	30
3	20
4	30
5	30
6	50
7	0
8	20
9	10
10	0

[그림 8-6 쿼리 실행 결과]

연령대(10세 단위)가 잘 나뉘었음을 확인할 수 있다. 이제 연령별로 탑승객 수와 생존자 수, 생존율을 구해 보자.

```
SELECT FLOOR(AGE/10)*10 AGEBOARD,  
COUNT(PASSENGERID) N_PASSENGERS,  
SUM(SURVIVED) N_SURVIVED,  
SUM(SURVIVED)/COUNT(PASSENGERID) SURVIVED_RATE  
FROM MYDATA.DATASET4  
GROUP  
BY 1  
;
```

	AGEBOARD	N_PASSENGERS	N_SURVIVED	SURVIVED_RATE
1	20	220	77	0.35
2	30	167	73	0.437125748502994
3	50	48	20	0.4166666666666667
4	0	62	38	0.612903225806452
5	10	102	41	0.401960784313726
6	40	89	34	0.382022471910112
7	60	19	6	0.315789473684211
8	70	6	0	0
9	80	1	1	1

[그림 8-7 쿼리 실행 결과]

연령대로 오름차순해 결과를 보기 좋게 정리해 보자.

```
SELECT FLOOR(AGE/10)*10 AGEBOARD,  
COUNT(PASSENGERID) N_PASSENGERS,  
SUM(SURVIVED) N_SURVIVED,  
SUM(SURVIVED)/COUNT(PASSENGERID) SURVIVED_RATE  
FROM MYDATA.DATASET4  
GROUP  
BY 1  
ORDER  
BY 1  
;
```

	<b>AGEBAND</b>	<b>N_PASSENGERS</b>	<b>N_SURVIVED</b>	<b>SURVIVED_RATE</b>
1	0	62	38	0.612903225806452
2	10	102	41	0.401960784313726
3	20	220	77	0.35
4	30	167	73	0.437125748502994
5	40	89	34	0.382022471910112
6	50	48	20	0.4166666666666667
7	60	19	6	0.315789473684211
8	70	6	0	0
9	80	1	1	1

[그림 8-8 쿼리 실행 결과]

결과를 보면 20대 탑승객 수가 가장 많았고, 70대를 제외하면 60대의 생존율이 가장 낮았다. 생존율이 가장 높았던 그룹은 0~9세 아동으로 나타난다.

이제 연령에 성별을 추가해 좀 더 세부적으로 생존율을 살펴보자.

```

SELECT FLOOR(AGE/10)*10 AGEBAND,
SEX,
COUNT(PASSENGERID) N_PASSENGERS,
SUM(SURVIVED) N_SURVIVED,
SUM(SURVIVED)/COUNT(PASSENGERID) SURVIVED_RATE
FROM MYDATA.DATASET4
GROUP
BY 1,2
ORDER
BY 2,1
;

```

	<b>AGEBAND</b>	<b>SEX</b>	<b>N_PASSENGERS</b>	<b>N_SURVIVED</b>	<b>SURVIVED_RATE</b>
1	0	female	30	19	0.6333333
2	10	female	45	34	0.7555556
3	20	female	72	52	0.7222222
4	30	female	60	50	0.8333333
5	40	female	32	22	0.6875000
6	50	female	18	16	0.8888889
7	60	female	4	4	1.0000000
8	0	male	32	19	0.5937500
9	10	male	57	7	0.1228070
10	20	male	148	25	0.1689189

[그림 8-9 쿼리 실행 결과]

결과를 보면 50대 여성의 생존율이 가장 높게 나타나고, 10대 남성의 생존율이 가장 낮게 나타난다(70대 제외). 만약 남성, 여성의 동일 연령대별 생존율 차이를 비교하려면 어떻게 해야 할까?

남성, 여성의 동일 연령대별 생존율 차이를 구하려면, 위 테이블을 2개(남성, 여성)로 나눈 뒤 연령대로 2개 테이블을 조인하면 구하려는 값을 계산할 수 있다.

먼저 위 결과를 2개 테이블(남성, 여성)로 구분해 보자.

```
SELECT FLOOR(AGE/10)*10 AGEBAND,  
SEX,  
COUNT(PASSENGERID) N_PASSENGERS,  
SUM(SURVIVED) N_SURVIVED,  
SUM(SURVIVED)/COUNT(PASSENGERID) SURVIVED_RATE  
FROM MYDATA.DATASET4  
GROUP  
BY 1,2  
HAVING SEX = 'male'  
;
```

AGEBAND	SEX	N_PASSENGERS	N_SURVIVED	SURVIVED_RATE	
1	20	male	148	25	0.1689189
2	30	male	107	23	0.2149533
3	50	male	30	4	0.1333333
4	0	male	32	19	0.5937500
5	10	male	57	7	0.1228070
6	40	male	57	12	0.2105263
7	60	male	15	2	0.1333333
8	70	male	6	0	0.0000000
9	80	male	1	1	1.0000000

[그림 8-10 쿼리 실행 결과]

```

SELECT FLOOR(AGE/10)*10 AGEBAND,
SEX,
COUNT(PASSENGERID) N_PASSENGERS,
SUM(SURVIVED) N_SURVIVED,
SUM(SURVIVED)/COUNT(PASSENGERID) SURVIVED_RATE
FROM MYDATA.DATASET4
GROUP
BY 1,2
HAVING SEX = 'female'
;

```

	AGEBAND	SEX	N_PASSENGERS	N_SURVIVED	SURVIVED_RATE
1	30	female	60	50	0.8333333
2	20	female	72	52	0.7222222
3	10	female	45	34	0.7555556
4	0	female	30	19	0.6333333
5	50	female	18	16	0.8888889
6	40	female	32	22	0.6875000
7	60	female	4	4	1.0000000

[그림 8-11 쿼리 실행 결과]

이제 2개 테이블을 조인하면, 성별 동일 연령대별 생존율 차이를 비교해 볼 수 있다. 2개 테이블을 서브 쿼리로 생성하고 조인해 보자.

```

SELECT A.AGEBAND,
A.SURVIVED_RATE MALE_SURVIVED_RATE,
B.SURVIVED_RATE FEMALE_SURVIVED_RATE,
B.SURVIVED_RATE - A.SURVIVED_RATE SURVIVED_RATE_DIFF
FROM
(SELECT FLOOR(AGE/10)*10 AGEBAND,
SEX,
COUNT(PASSENGERID) N_PASSENGERS,
SUM(SURVIVED) N_SURVIVED,
SUM(SURVIVED)/COUNT(PASSENGERID) SURVIVED_RATE
FROM MYDATA.DATASET4
GROUP
BY 1,2
HAVING SEX = 'male') A

```

```

LEFT
JOIN
(SELECT FLOOR(AGE/10)*10 AGEBAND,
SEX,
COUNT(PASSENGERID) N_PASSENGERS,
SUM(SURVIVED) N_SURVIVED,
SUM(SURVIVED)/COUNT(PASSENGERID) SURVIVED_RATE
FROM MYDATA.DATASET4
GROUP
BY 1,2
HAVING SEX = 'female') B
ON A.AGEBAND = B.AGEBAND
ORDER
BY A.AGEBAND
;

```

AGEBAND	MALE_SURVIVED RATE	FEMALE_SURVIVED RATE	SURVIVED RATE DIFF
1 0	0.59375	0.6333333333333333	0.0395833333333333
2 10	0.12280701754386	0.755555555555556	0.632748538011696
3 20	0.168918918918919	0.722222222222222	0.553303303303303
4 30	0.214953271028037	0.833333333333333	0.618380062305296
5 40	0.210526315789474	0.6875	0.476973684210526
6 50	0.133333333333333	0.888888888888889	0.755555555555556
7 60	0.133333333333333	1	0.8666666666666667
8 70	0	NA	NA
9 80	1	NA	NA

[그림 8-12 쿼리 실행 결과]

SURVIVED\_RATE\_DIFF는 연령별 여성의 생존율 - 남성의 생존율을 나타낸다. 전반적으로 여성의 경우 모든 연령대에서 60% 이상의 생존율을 보인다. 반대로 남성의 경우 10, 20대의 생존율이 50, 60대와 비슷하게 나타난다.

성별과 연령에 따라 생존율이 다른 이유는 무엇일까?

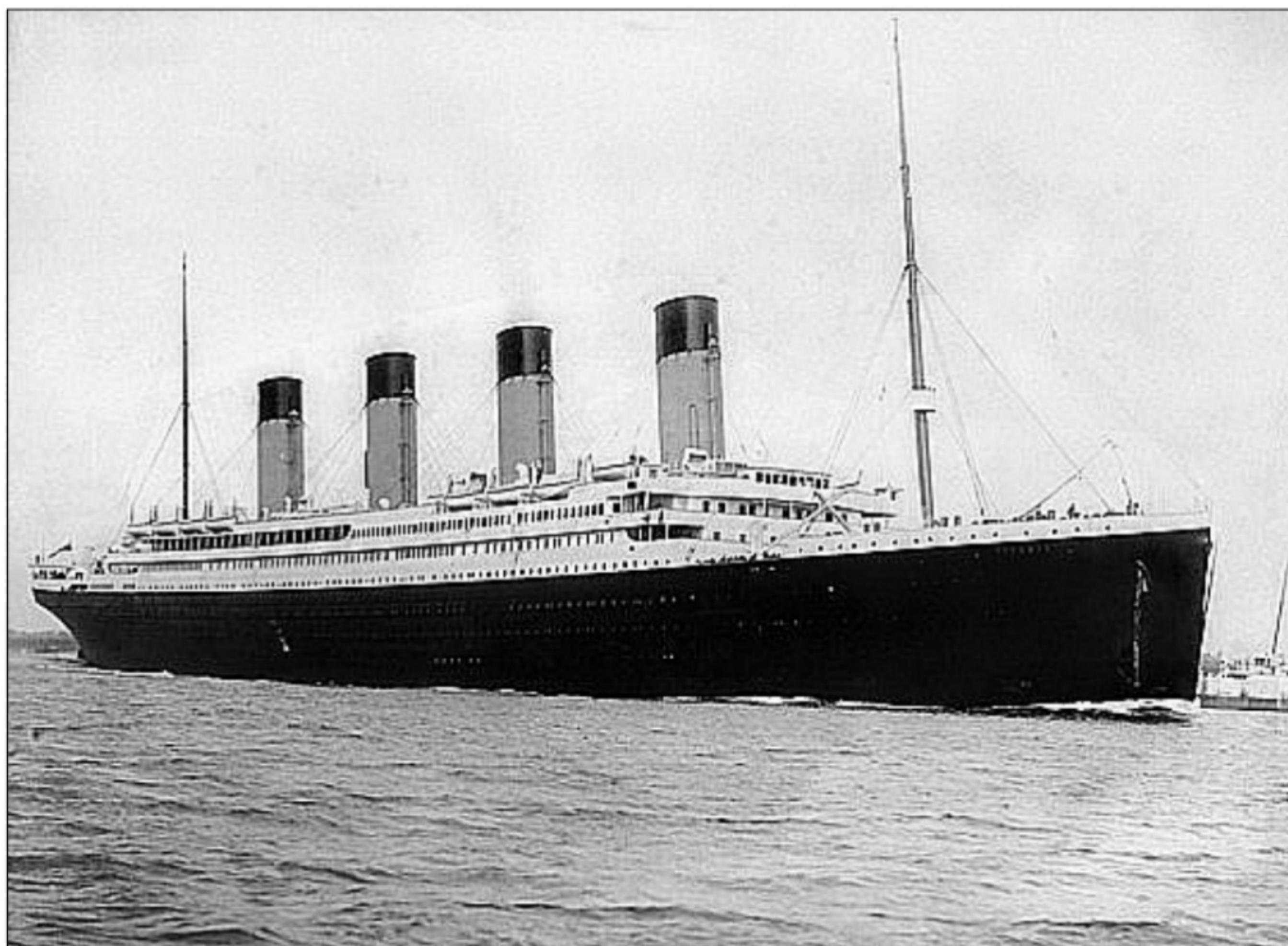
스웨덴 유플라 대학 경제학 교수인 미카엘 엘린더와 오스카 에릭손은 타이타닉처럼 배가 침몰할 때 선장과 승무원, 승객의 생존율을 비교한 연구 결과, 지난 1852년부터 2011년까지 세계 30개국에서 일어난 해상 사고를 분석한 조사에서 사고 시 가장 생존율이 높은 사람들은 다름 아닌 선장과 승무원으로 드러났고, ‘여성과 어린이 먼저’라는 기사도 대신 ‘모든 사람이 자신만 생각’하는 패턴을 보

였다고 주장했다. (<https://news.joins.com/article/14492475>)

위 연구 결과에 따르면 가장 생존율이 높은 집단은 선장과 승무원이라고 한다. 즉 사고가 발생하면, 인간은 자신의 안전을 가장 우선시한다는 것이다. 그렇다면 타이타닉 사고에서는 왜 특이하게 성별, 연령에 따라 생존율에 차이가 있었을까?

당시 타이타닉호 선장이었던 에드워드 존 스미스는 승객 중에서 어린이, 여자, 남자 순으로 탈출도록 했고, 총으로 공포를 쏘면서 이성을 잃은 사람들이 질서를 유지하도록 하게 했으며, 배와 운명을 함께하는 직업의식과 책임감을 보였다. (<https://news.joins.com/article/14492475>)

선장이 구조 활동에서 어린이, 여자, 남자 순으로 탈출시켜 연령과 성별에 따라 생존율에 차이가 존재했던 것으로 보인다.



[그림 8-13 타이타닉호]

만약 이런 사실이 알려지지 않았다면, 타이타닉 구조 활동의 결과는 영국식 기사도의 결과라는 잘못된 주장의 근거가 될 수도 있었을 것이다. 이처럼 데이터는 어떻게 해석하느냐에 따라 잘못된 주장의 근거로 사용될 수도 있다. 따라서 데이터

를 분석할 때는 한 가지 관점이 아닌 다양한 관점과 가능성을 열어 두고 해석하는 것이 중요하다.

### 3) Pclass(객실 등급)

이제 생존율과 객실 등급 사이의 관계를 살펴보려고 한다. 먼저 객실 등급에는 어떤 값이 있는지 살펴보자.

```
SELECT DISTINCT PCCLASS  
FROM MYDATA.DATASET4  
;
```

PCCLASS	
1	3
2	1
3	2

[그림 8-14 Pclass 종류]

타이타닉호의 객실 등급은 1등실, 2등실, 3등실로 나뉘어 있었다.

#### 1등실

1등실에는 총 329명의 부유한 승객이 타고 있었다. 급한 사정보다는 여유를 즐기려고 승선한 승객이 많았다. 객실은 최상층인 보트 갑판부터 갑판 E(상갑판) 까지 설치되어 있었으며, 호화 호텔 수준이었고, 개인 목욕탕이 있었다.

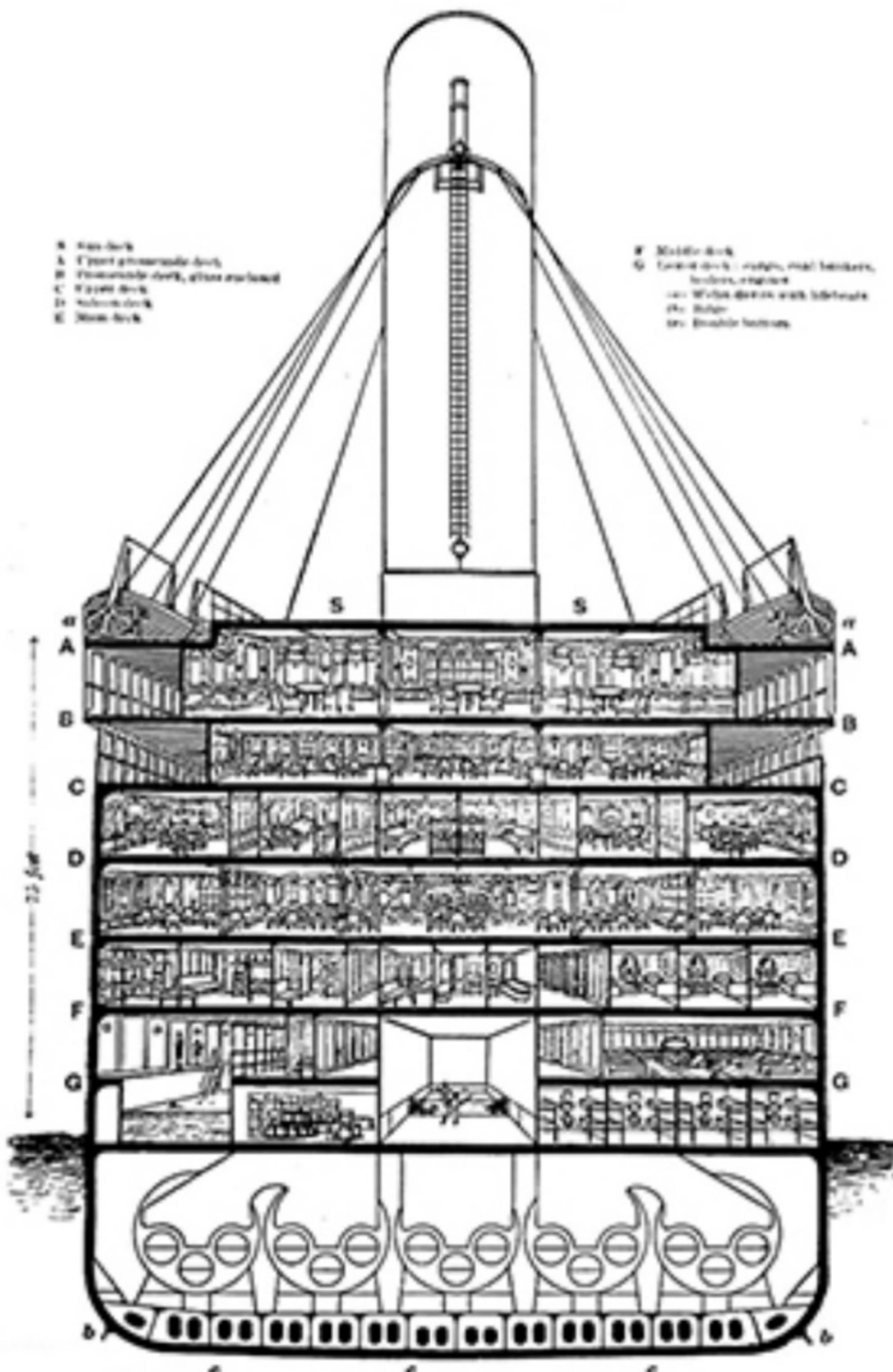
#### 2등실

2등실에는 총 285명의 중산층 승객이 타고 있었다. 전체적으로 봤을 때 1등실만큼 좋지는 않았지만, 그래도 비교적 편리한 시설이 설치되어 있었다. 객실은 갑판 D부터 갑판 G까지 설치되어 있었다. 흡연실(갑판 B), 레스토랑(갑판 B), 도서관(갑판 C), 상점 등이 있었다.

### 3등실

3등실에는 총 710명의 가난한 승객들이 타고 있었다. 주로 아메리칸 드림으로 미국에서 새로운 보금자리를 얻기 위해 승선한 승객들이었다. 객실은 2등실과 마찬가지로 갑판 D부터 갑판 G까지 설치되어 있었다. 시설은 1등실과 2등실만 못하고, 엔진이 가동되는 소리가 울려 퍼졌으나 다른 배들보다 비교적 좋은 대우를 해 주었다. 배에 탑승하기 전에는 검역을 걸쳐서 전염병이나 이/벼룩을 확인했고, 여자와 남자는 배의 앞머리와 뒷머리에 각각 따로 떨어져 승선했으나 가족 단위일 경우 같이 승선할 수 있었다.

([https://ko.wikipedia.org/wiki/RMS\\_Titanic](https://ko.wikipedia.org/wiki/RMS_Titanic))



[그림 8-15 타이타닉 구조도]

객실 등급별로 승객 수와 생존자 수, 생존율을 계산해 보자.

```
SELECT PCLASS,
       COUNT(PASSENGERID) N_PASSENGERS,
       SUM(SURVIVED) N_SURVIVED,
       SUM(SURVIVED)/COUNT(PASSENGERID) SURVIVED_RATE
  FROM MYDATA.DATASET4
 GROUP
   BY PCLASS
 ORDER
   BY 1
;
```

PCLASS	N_PASSENGERS	N_SURVIVED	SURVIVED_RATE
1	186	122	0.655913978494624
2	173	83	0.479768786127168
3	355	85	0.23943661971831

[그림 8-16 쿼리 실행 결과]

실제 기록과 추출한 결과를 비교해 보면, 탑승객 수에 차이가 존재한다. 당시에는 탑승이 전산으로 처리되지 않았을 것이므로, 일부 승객의 정보가 유실될 가능성이 있다.

생존율은 1등석, 2등석, 3등석 순으로 높게 나타난다. 앞에서 성별, 연령에 따라 생존율이 다른 점은 선장의 구조 정책 때문이었다. 분명히 어린이-여성-남성 순서로 구조되었다면, 객실 등급의 생존율에 이렇게 큰 차이가 있을 수 없을 것이다.

각 객실의 위치를 보면, 이 결과를 어느 정도 이해할 수 있다. 먼저 상위 등급의 객실일수록 배의 상층에 위치한다. 즉 사고가 발생했을 때 탈출하기에 더 용이했을 것이다.

추가로 객실 등급과 연령, 성별을 조합해 생존율을 살펴보자.

```

SELECT PCLASS,
SEX,
COUNT(PASSENGERID) N_PASSENGERS,
SUM(SURVIVED) N_SURVIVED,
SUM(SURVIVED)/COUNT(PASSENGERID) SURVIVED_RATE
FROM MYDATA.DATASET4
GROUP
BY PCLASS,SEX
ORDER
BY 2,1
;

```

PCLASS	SEX	N_PASSENGERS	N_SURVIVED	SURVIVED_RATE
1	1	female	85	82
2	2	female	74	68
3	3	female	102	47
4	1	male	101	40
5	2	male	99	15
6	3	male	253	38

[그림 8-17 쿼리 실행 결과]

앞에서 살펴본 것처럼 여성의 생존율이 높게 나타난다. 3등석 여성의 생존율이 1등석 남성의 생존율보다 높은 것은 실제로 객실 등급보다 연령, 성별이 구조에서 우선 고려된 것으로 보인다. 추가로 연령도 함께 고려해 생존율을 살펴보자.

```

SELECT PCLASS,
SEX,
FLOOR(AGE/10)*10 AGEBAND,
COUNT(PASSENGERID) N_PASSENGERS,
SUM(SURVIVED) N_SURVIVED,
SUM(SURVIVED)/COUNT(PASSENGERID) SURVIVED_RATE
FROM MYDATA.DATASET4
GROUP
BY PCLASS,SEX, FLOOR(AGE/10)*10
ORDER
BY 2,1
;

```

PCLASS	SEX	AGEBAND	N_PASSENGERS	N_SURVIVED	SURVIVED_RATE
1	1	female	30	27	1.0000000
2	1	female	50	12	0.9166667
3	1	female	10	13	1.0000000
4	1	female	40	13	1.0000000
5	1	female	60	3	1.0000000
6	1	female	20	16	0.9375000
7	1	female	0	1	0.0000000
8	2	female	20	25	0.8800000
9	2	female	10	8	1.0000000
10	2	female	50	6	0.8333333

[그림 8-18 쿼리 실행 결과]

앞에서 살펴본 것처럼 유아일수록 생존율이 높았고, 모든 객실 등급에서 여성의 생존율이 더 높게 나타난다.

타이타닉호가 침몰할 때 승무원과 각 호실 또는 남녀, 아동 간 구조된 인원과 구조되지 못한 인원들은 다음과 같다. 1등실에 탑승한 어린이 중에서는 당시 2살 이던 헬렌 로레인 알리슨 양만 유일하게 구조되지 못했으며, 2등실에 탑승한 어린이는 전원 구조되었다. 반면에 2등실에 탑승한 성인 남자 승객은 168명 중 겨우 14명만 구조되어 92%에 달하는 인원인 154명이 바다에 수장되었다. 부자들과 사회 각 계층의 고위 인사들은 1등실에 탑승해 상당수가 구조되었지만, 3등 실에 탑승한 이민자들은 과반수가 구조되지 못했는데, 특이하게 이민자들보다도 신분이 낮은 하인들이 많이 구조되었다. 하인들의 경우 고위인사들을 보좌하도록 1등실에 탑승한 상태였는데, 해당 하인의 주인들이 자신의 하인을 구조하라고 승무원들에게 요구했기 때문이었다. 3등실의 생존율이 가장 낮은 이유는 가장 낮은 구역에 위치할 뿐만 아니라 여러 구역이 잠겨 있었고, 배에 친숙하지 않은 승객들이 길을 혜택받기 때문일 가능성이 크다. 확인되지는 않았지만, 승무원들이 이들을 통제했을 가능성도 있다.

([https://ko.wikipedia.org/wiki/RMS\\_Titanic](https://ko.wikipedia.org/wiki/RMS_Titanic))

### 3 EMBARKED

다음으로 승선 항구에 따른 승전자 수와 생존율의 관계를 살펴보자. 앞에서 살펴본 내용과 유사한 내용이므로 쿼리에 대한 자세한 설명은 생략한다.

#### 1) 승선 항구별 승객 수

```
SELECT EMBARKED,
       COUNT(PASSENGERID) N_PASSENGERS
    FROM MYDATA.DATASET4
   GROUP
      BY 1
   ORDER
      BY 1
;
```

EMBARKED		N_PASSENGERS
1		2
2	C	130
3	Q	28
4	S	554

[그림 8-19 쿼리 실행 결과]

#### 2) 승선 항구별, 성별 승객 수

```
SELECT EMBARKED,
       SEX,
       COUNT(PASSENGERID) N_PASSENGERS
    FROM MYDATA.DATASET4
   GROUP
      BY 1,2
   ORDER
      BY 1,2
;
```

	<b>EMBARKED</b>	<b>SEX</b>	<b>N_PASSENGERS</b>
1		female	2
2	C	female	61
3	C	male	69
4	Q	female	12
5	Q	male	16
6	S	female	186
7	S	male	368

[그림 8-20 쿼리 실행 결과]

### 3) 승선 항구별, 성별 승객 비중(%)

각 승선 항구별 여성, 남성 승객 수를 계산해 보았다. 추가로 여성 승객, 남성 승객의 비중이 얼마나 되는지 계산해 보자. 항구별 전체 승객 수를 계산하고, 남성, 여성 승객 수를 전체 승객 수로 나누면, 남성, 여성의 승객 비중(%)을 계산할 수 있다.

#### a) 승선 항구별 전체 승객 수

```
SELECT EMBARKED,
COUNT(PASSENGERID) N_PASSENGERS
FROM MYDATA.DATASET4
GROUP
BY 1
;
```

#### b) 승선 항구별, 성별 승객 수

```
SELECT EMBARKED,
SEX,
COUNT(PASSENGERID) N_PASSENGERS
FROM MYDATA.DATASET4
GROUP
BY 1,2
;
```

### c) 테이블 결합

```
SELECT A.EMBARKED,
A.SEX,
A.N_PASSENGERS,
B.N_PASSENGERS N_PASSENGERS_TOT,
A.N_PASSENGERS/B.N_PASSENGERS PASSENGERS_RAT
FROM
(SELECT EMBARKED,
SEX,
COUNT(PASSENGERID) N_PASSENGERS
FROM MYDATA.DATASET4
GROUP
BY 1,2) A
LEFT
JOIN
(SELECT EMBARKED,
COUNT(PASSENGERID) N_PASSENGERS
FROM MYDATA.DATASET4
GROUP
BY 1) B
ON A.EMBARKED = B.EMBARKED
```

과정이 다소 복잡하지만, 테이블 결합은 쿼리 작성에서 매우 중요한 부분이므로 많이 연습해 보자.

## 4 탑승객 분석

앞에서 각 요인과 생존 여부의 관계를 살펴보았다. 이번 챕터에서는 타이타닉호에 승선한 승객들을 분석해 보려고 한다.

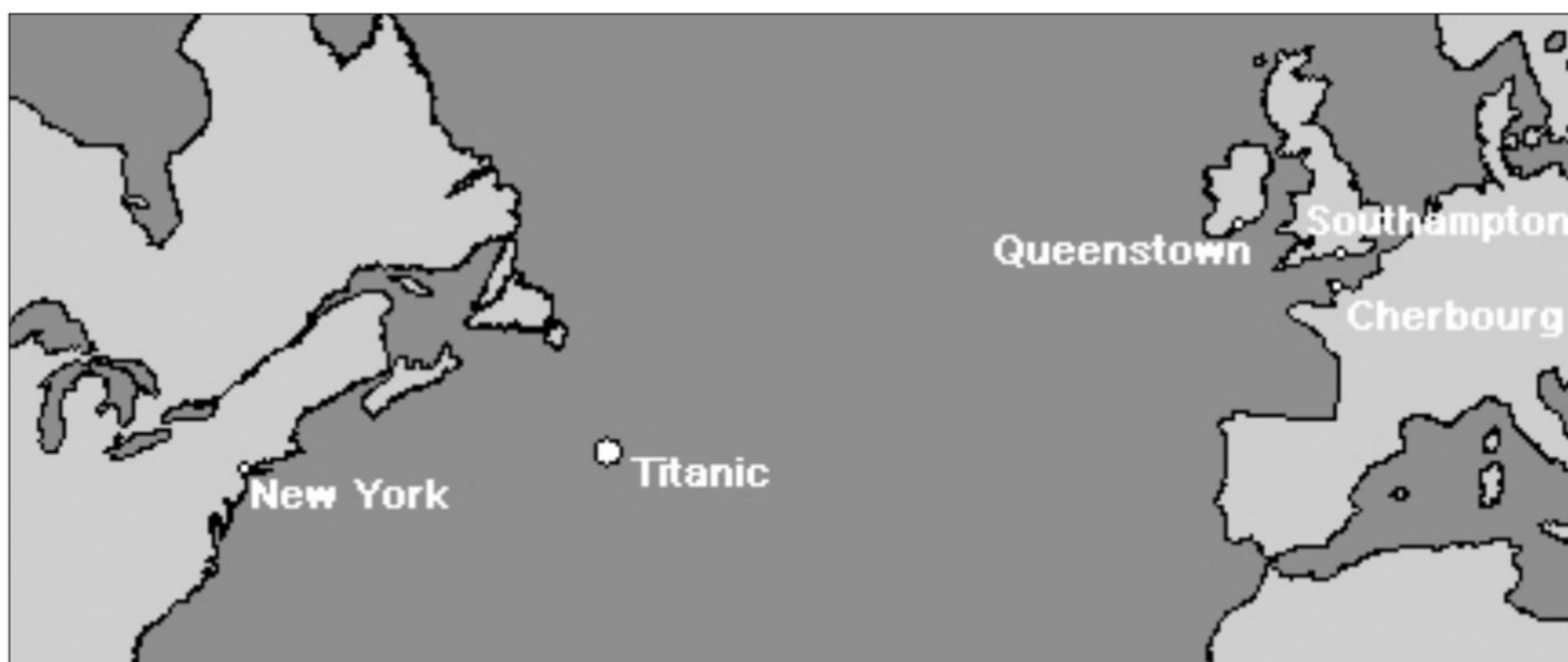
### 1) 출발지, 도착지별 승객 수

칼럼 중 Boarded, Destination이라는 칼럼이 존재한다. 2개 칼럼은 출발지와 목적지를 의미한다. 출발지와 목적지별 승객 수를 구해 보자.

```
SELECT BOARDED,  
DESTINATION,  
COUNT(PASSENGERID) N_PASSENGERS  
FROM MYDATA.DATASET4  
GROUP  
BY BOARDED,  
DESTINATION  
ORDER  
BY 3 DESC
```

BOARDED	DESTINATION	N_PASSENGERS
1 Southampton	New York City	90
2 Southampton	New York, New York, US	39
3 Cherbourg	New York, New York, US	33
4 Southampton	Chicago, Illinois, US	33
5 Southampton	Detroit, Michigan, US	19
6 Southampton	Montreal, Quebec, Canada	18
7 Southampton	Winnipeg, Manitoba, Canada	16
8 Queenstown	New York City	10
9 Southampton	Brooklyn, New York, US	9
10 Cherbourg	Ottawa, Ontario, Canada	8

[그림 8-21 쿼리 실행 결과]



[그림 8-22 타이타닉 경로 및 침몰 지역]

타이타닉호는 지도 우측 상단에 위치한 영국의 Southampton에서 출발해 New York으로 향했다고 한다. Southampton에서 탑승해 New York으로 향하는 승객이 가장 많았고, 다음으로는 Cherbourg에서 출발해 New York으로 향하는 승객이 많았다.

만약 상위 5개 경로를 선택한 승객들의 이름을 추출하려면 어떻게 해야 할까?

먼저 탑승객 수로 순위를 매긴다. 탑승객 수가 가장 많은 경로가 1이 되어야 하므로, 내림차순(DESC)으로 정렬한다.

```
SELECT *,  
ROW_NUMBER() OVER(ORDER BY N_PASSENGERS DESC) RNK  
FROM  
(SELECT BOARDED,  
DESTINATION,  
COUNT(PASSENGERID) N_PASSENGERS  
FROM MYDATA.DATASET4  
GROUP  
BY BOARDED,  
DESTINATION) BASE  
;
```

BOARDED	DESTINATION	N_PASSENGERS	RNK
1 Southampton	New York City	90	1
2 Southampton	New York, New York, US	39	2
3 Cherbourg	New York, New York, US	33	3
4 Southampton	Chicago, Illinois, US	33	4
5 Southampton	Detroit, Michigan, US	19	5
6 Southampton	Montreal, Quebec, Canada	18	6
7 Southampton	Winnipeg, Manitoba, Canada	16	7
8 Queenstown	New York City	10	8
9 Southampton	Brooklyn, New York, US	9	9
10 Cherbourg	Ottawa, Ontario, Canada	8	10

[그림 8-23 쿼리 실행 결과]

다음으로 상위 5개 경로를 선택한다. 이후 해당 경로를 테이블로 생성한다.

```

CREATE TEMPORARY TABLE MYDATA.ROUTE AS
SELECT BOARDED,
DESTINATION
FROM
(SELECT *,  

ROW_NUMBER() OVER(ORDER BY N_PASSENGERS DESC) RNK
FROM
(SELECT BOARDED,  

DESTINATION,  

COUNT(PASSENGERID) N_PASSENGERS
FROM MYDATA.DATASET4
GROUP  

BY BOARDED,  

DESTINATION) BASE) BASE
WHERE RNK BETWEEN 1 AND 5
;

```

마지막으로 생성한 경로에 해당하는 승객들의 이름을 추출한다.

```

SELECT NAME_WIKI,  

A.BOARDER,  

A.DESTINATION
FROM MYDATA.DATASET4 A
INNER
JOIN MYDATA.ROUTE B
ON A.BOARDER = B.BOARDER AND A.DESTINATION = B.DESTINATION

```

	NAME_WIKI	BOARDED	DESTINATION
1	Cumings, Mrs. Florence Briggs (n.e Thayer)	Cherbourg	New York, New York, US
2	Heikkinen, Miss Laina	Southampton	New York City
3	Allen, Mr. William Henry	Southampton	New York City
4	P.Isson, Master G.sta Leonard	Southampton	Chicago, Illinois, US
5	Saundercock, Mr. William Henry	Southampton	New York City
6	P.Isson, Miss Torborg Danira	Southampton	Chicago, Illinois, US
7	Meyer, Mr. Edgar Joseph	Cherbourg	New York, New York, US
8	Holverson, Mr. Alexander Oskar	Southampton	New York, New York, US
9	Cann, Mr. Ernest Charles	Southampton	New York City
10	Harper, Mrs. Myna (n.e Haxtun)	Cherbourg	New York, New York, US

[그림 8-24 쿼리 실행 결과]

## 2) Hometown별 탑승객 수 및 생존율

o]제 Hometown별 탑승객 수와 탑승객 대비 생존율을 구해 보자.

```
SELECT HOMETOWN,
SUM(1) N_PASSENGERS,
SUM(SURVIVED)/SUM(1) SURVIVED_RATIO
FROM MYDATA.DATASET4
GROUP
BY 1
;
```

	HOMETOWN	N_PASSENGERS	SURVIVED_RATIO
1	Bridgerule, Devon, England	2	0.0000000
2	New York, New York, US	44	0.7727273
3	Jyv.skyl., Finland	1	1.0000000
4	Scituate, Massachusetts, US	2	0.5000000
5	Birmingham, West Midlands, England	1	0.0000000
6	Dorchester, Massachusetts, US	1	0.0000000
7	Bjuv, Sk.ne, Sweden	4	0.0000000
8	St. Charles, Illinois, US	3	1.0000000
9	Zahl., Lebanon, Ottoman Empire	1	1.0000000
10	Motala, .sterg.tland, Sweden	2	1.0000000

[그림 8-25 쿼리 실행 결과]

탑승객 수가 10명 이상이면서 생존율이 0.5 이상인 HOMETOWN을 출력하려면 어떻게 해야 할까? HAVING을 사용하면 GROUPING한 데이터에 조건을 만들 수 있다. 다음과 같이 처리해 보자.

```
SELECT HOMETOWN,  
SUM(1) N_PASSENGERS,  
SUM(SURVIVED)/SUM(1) SURVIVED_RATIO  
FROM MYDATA.DATASET4  
GROUP  
BY 1  
HAVING SUM(SURVIVED)/SUM(1) >= 0.5 AND SUM(1) >= 10  
;
```

	HOMETOWN	N_PASSENGERS	SURVIVED_RATIO
1	New York, New York, US	44	0.7727273
2	London, England, UK	19	0.6842105
3	Paris, France	13	0.5384615

[그림 8-26 쿼리 실행 결과]

## 5 상관 분석(Correlation Analysis)

앞에서 생존 여부와 각 요인의 관계를 살펴보았다. 타이타닉 데이터의 경우 19개의 변수를 가지고 있다. 변수별로 생존 여부와의 관계를 모두 분석하는 것은 쉬운 일이 아니다.

그리고 쿼리를 통해 분석한 결과로는 변수 사이의 상관관계(양의 상관관계, 음의 상관관계)를 파악할 수 없고, 상관도를 숫자로 측정할 수 없다. 그렇기 때문에 통상적으로 변수 사이의 관계를 파악하려면 Python, R 등의 분석 툴을 이용한다.

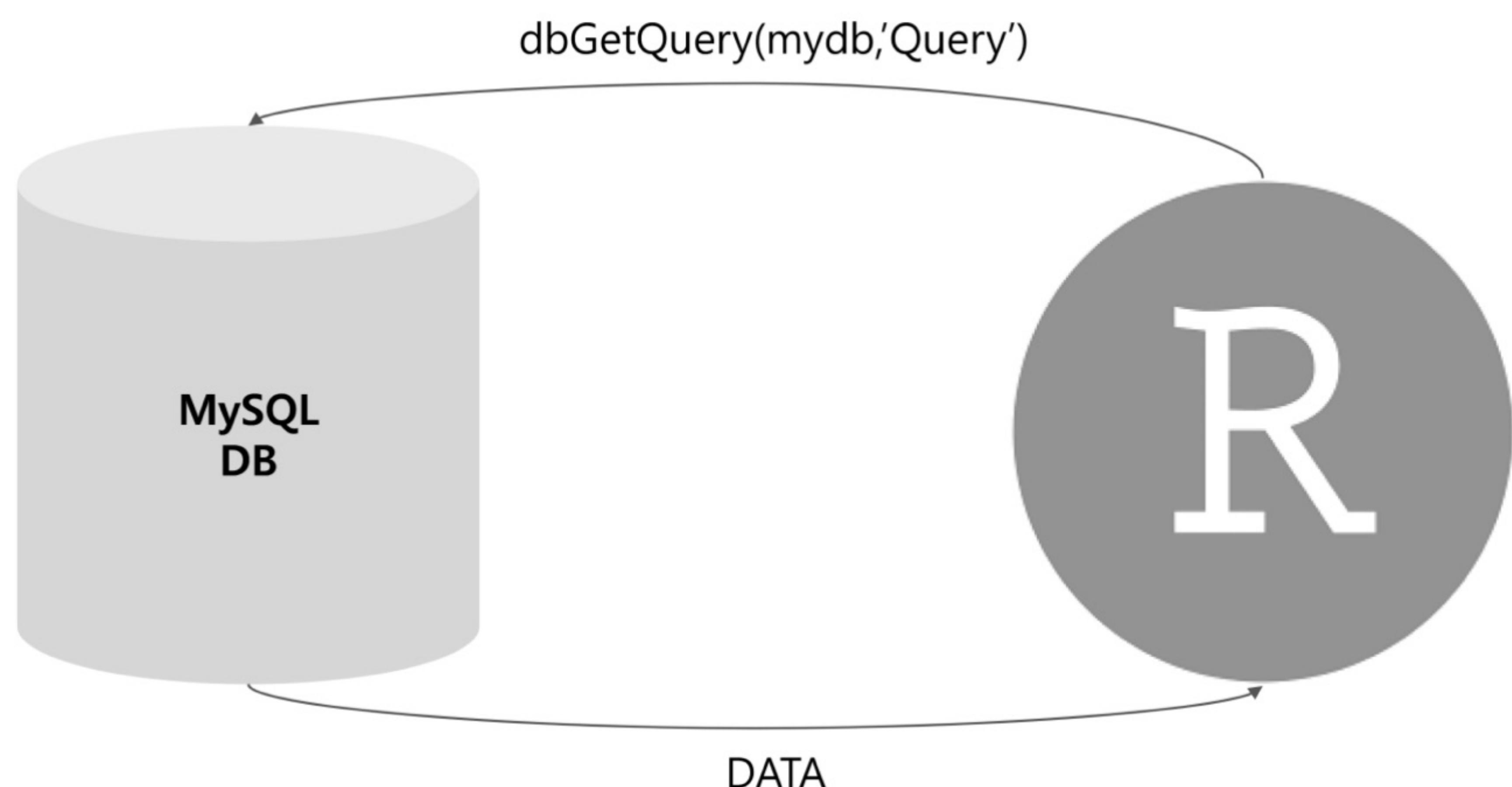
우리도 R을 이용해 상관관계를 분석하고, 인사이트를 찾아보자. 어떻게 SQL과 R이 분석에서 사용되는지에 집중해서 학습하자. R의 설치는 9장을 참고한다.

### 1) SQL로 데이터 불러오기

```
# MySQL과 R을 연동해 주는 패키지 설치  
install.packages('RMySQL') #1  
# 패키지 실행  
library(RMySQL) #2 (MySQL 접속 패키지)  
library(dplyr) # 데이터 전처리 패키지  
  
# 접속 정보 입력  
mydb = dbConnect(MySQL(),  
user='id', # MySQL 접속 계정  
password='password', # MySQL 접속 계정 패스워드  
dbname='mydata', # 데이터베이스명  
host='localhost') # DB IP (local에 접속하는 경우: localhost) #3  
# 데이터 data  
data = dbGetQuery(mydb, 'select * from mydata.dataset4') #4
```

먼저 R과 MySQL을 연결해 주는 패키지를 설치하고(#1), 패키지를 실행한다(#2). MySQL 접속 정보를 입력한다(#3). 쿼리를 작성하고, 쿼리의 결과를 data로 받아온다(#4).

dbGetQuery( ) 내부에 쿼리를 입력하고 실행해 MySQL DB를 R로 가져온다. 이렇게 가져온 데이터를 R에서 분석할 수 있다.



[그림 8-27 DB, R 연결 및 데이터 가져오기]

## 2) 문자열 숫자로 변경하기

상관 계수를 구하기 위해서 변수는 숫자 형태여야 한다. 분석을 위해 여성은 1, 남성은 0으로 표현한다.

```
data$Sex = ifelse(data$Sex == 'female',1,0)
```

### 3) 분석할 변수 가져오기

분석하려는 변수만 선택한다.

```
M = data %>% select(Survived, Age, SibSp, Parch, Fare, Sex)
```

### 4) 상관 계수 구하기

`cor(M)`

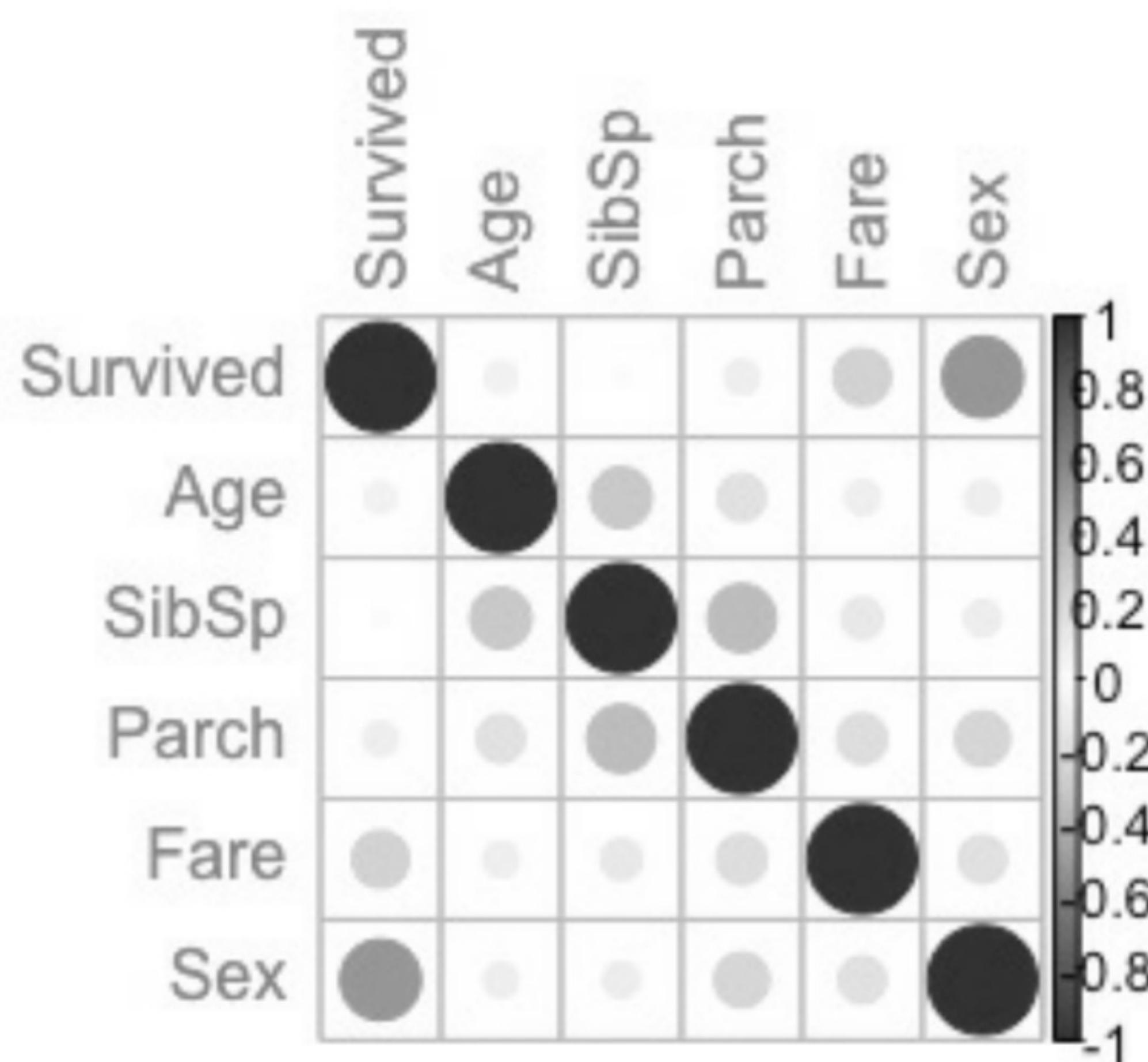
	Survived	Age	SibSp	Parch	Fare	Sex
Survived	1.00000000	-0.07722109	-0.01735836	0.09331701	0.26818862	0.53882559
Age	-0.07722109	1.00000000	-0.30824676	-0.18911926	0.09606669	-0.09325358
SibSp	-0.01735836	-0.30824676	1.00000000	0.38381986	0.13832879	0.10394968
Parch	0.09331701	-0.18911926	0.38381986	1.00000000	0.20511888	0.24697204
Fare	0.26818862	0.09606669	0.13832879	0.20511888	1.00000000	0.18499425
Sex	0.53882559	-0.09325358	0.10394968	0.24697204	0.18499425	1.00000000

[그림 8-28 Correlation Matrix]

변수별 상관 계수를 간단하게 구할 수 있다. 이제 결과를 시각화해 보자.

### 5) 상관관계 시각화하기

```
install.packages('corrplot') # 상관관계 분석 패키지 설치  
library(corrplot) # 상관관계 분석 패키지 불러오기  
  
corr = cor(M) # 상관 계수 계산하기  
corrplot(corr, method="circle") # 상관 계수 시각화
```



[그림 8-29 Correlation Plot]

원의 색상은 양과 음의 상관관계를 나타내고, 상관 계수의 절댓값이 클수록 원의 크기와 색상이 진하다. Survived와 각 변수의 관계를 보면, 성별이 증가할수록(여성일수록, 여성: 1, 남성: 0) 생존율이 높고, Age가 낮을수록 생존율이 높은 것으로 나타난다.

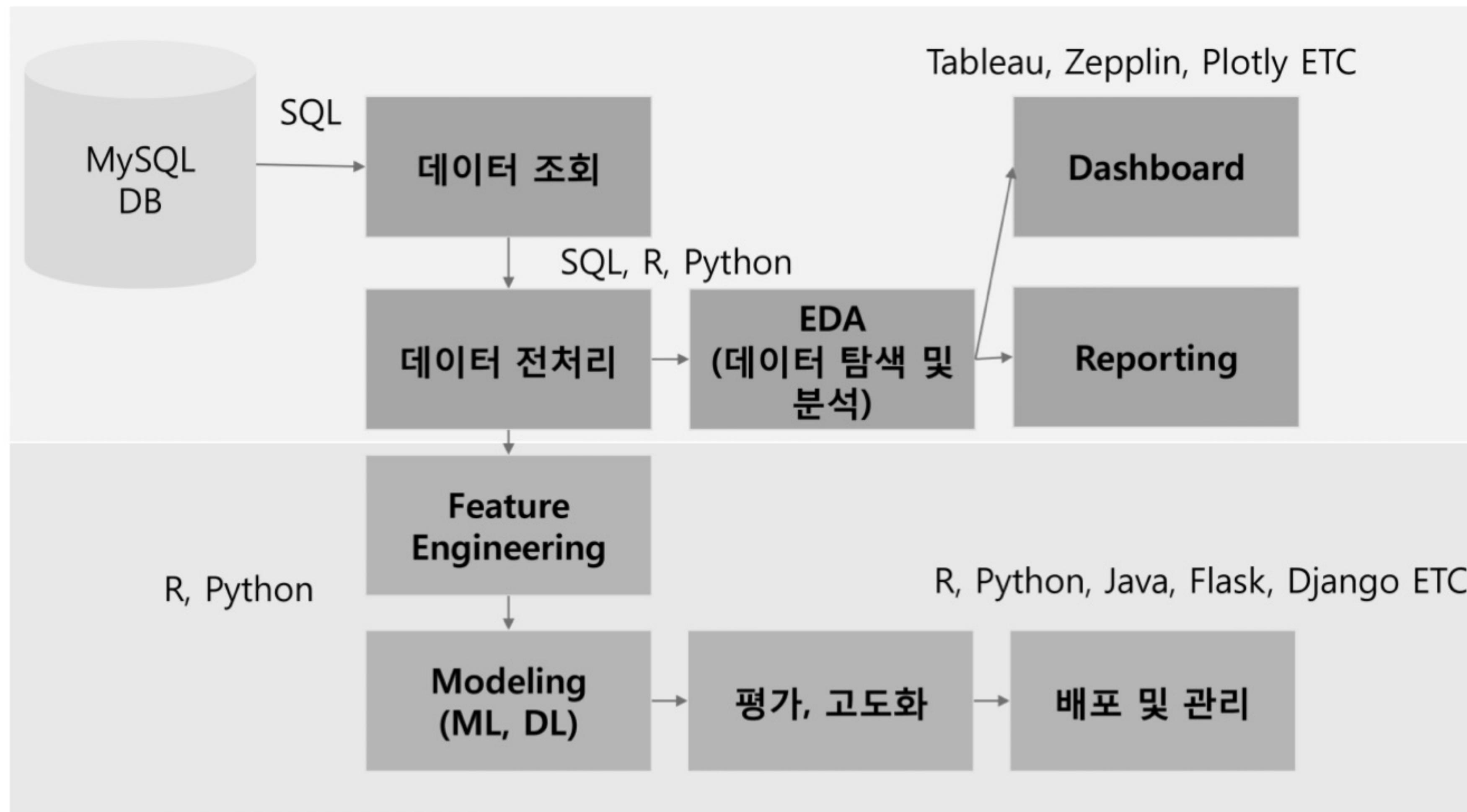
SibSp, Parch는 생존 여부와 큰 상관관계가 없어 보이고, Fare(티켓 가격)가 증가할수록 생존율이 높아지는 것으로 보인다.

우리가 SQL로 분석했던 내용과 같은 결과를 보이는데, 상관관계 분석을 통해서 각 변수 사이의 관계 정도를 측정할 수 있게 되었다. 생존 여부와 가장 밀접한 관계를 보이는 성별이었다.

다음 챕터에서는 R, Python과 같은 분석 툴을 통해서 타이타닉 데이터를 어떻게 분석할 수 있는지 살펴보자. 이 과정을 살펴보면 데이터 분석의 전반적인 과정을 이해하고, 추후 학습 방향에 도움이 될 수 있다.

## 6 데이터 분석 및 시각화

### 1) 업무별 데이터 분석 절차



[그림 8-30 Data Analyst, Data Scientist]

최근 데이터 분석 직군은 비즈니스 인사이트 발굴, 모니터링, 리포팅과 같은 업무를 담당하는 데이터 분석가와 머신러닝 모델 개발, 모델 배포 및 관리를 담당하는 ML Engineer / Data Scientist로 나뉜다.

두 가지 직군 모두 DB에 존재하는 데이터를 이용하게 되는데, 이때 SQL을 사용하게 된다. SQL을 이용해 DB에서 R, Python과 같은 분석 툴로 데이터를 임포트한 뒤 시각화, 통계 분석, 머신러닝과 같은 분석을 진행한다.

### 2) 분석 예제: 생존율 예측 모델 생성

머신러닝 모델 생성 과정을 간단하게 살펴보면서 SQL이 어떤 방식으로 활용되는지 파악해 보자.

### a) 데이터 임포트(SQL)

```
# 패키지 실행
library(RMySQL) # MySQL 접속 패키지
library(dplyr) # 데이터 전처리 패키지
library(ggplot2)

# 접속 정보 입력
mydb = dbConnect(MySQL(),
user='guest', # MySQL 접속 계정
password='guest', # MySQL 접속 계정 패스워드
dbname='mydata', # 데이터베이스명
host='localhost') # DB IP (localhost에 접속하는 경우: localhost)

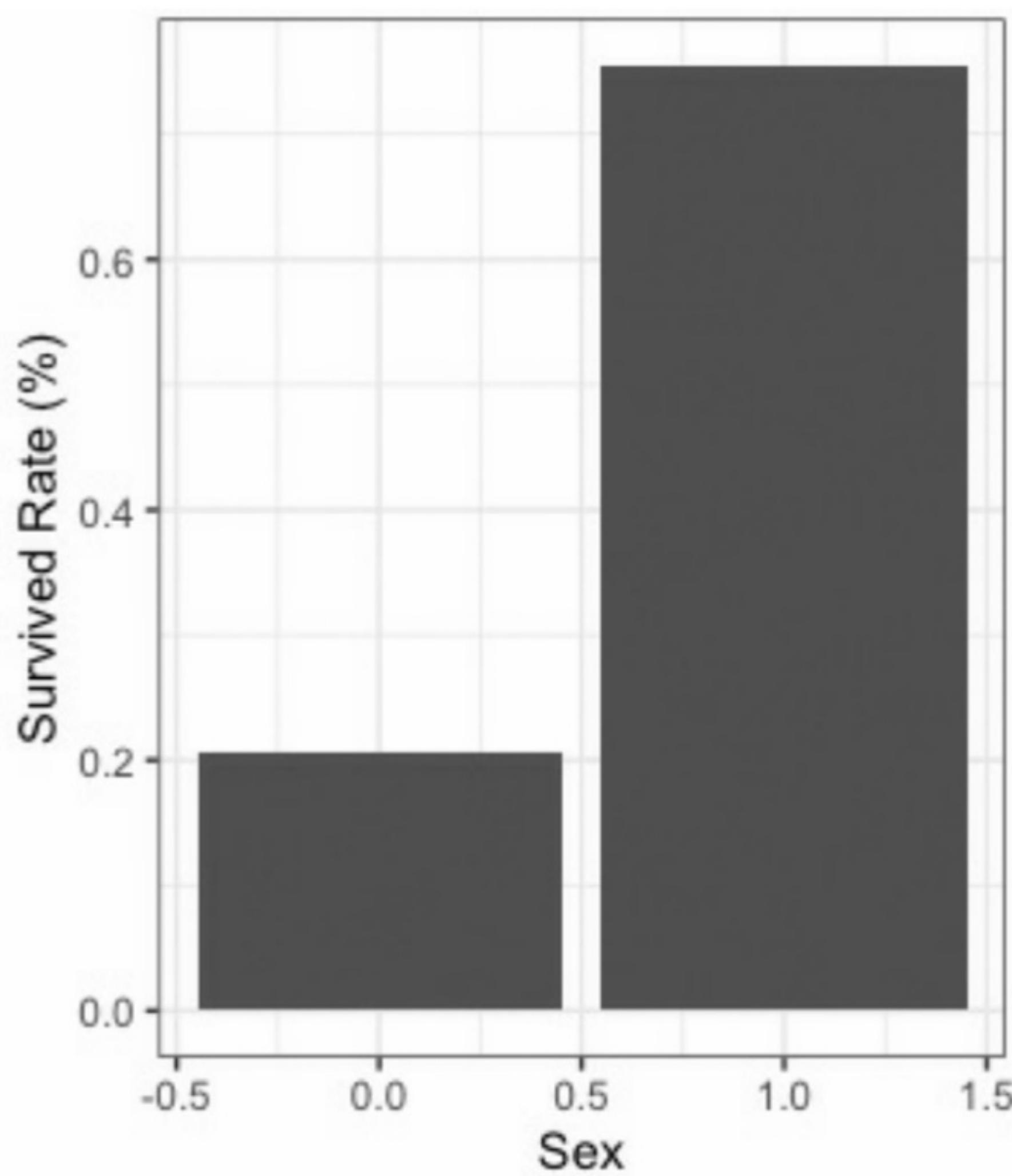
# 데이터 data
data = dbGetQuery(mydb, 'select * from mydata.dataset4')
```

앞에서 살펴본 것처럼 DB에 존재하는 타이타닉 데이터를 R로 임포트한다. 해당 예제는 분석이 이루어지는 절차에 더 집중하고자 한다. 사실 실무나 협업에서는 모든 정보가 테이블에 적재되지 않고 분산된 경우가 많다. 따라서 위의 쿼리보다 훨씬 복잡한 쿼리를 작성하게 된다.

### b) 생존율과 각 변수의 관계 파악 1(성별에 따른 생존율)

```
# 성별에 따른 생존율
stat = data %>% group_by(Sex) %>%
summarise(n_passengers = n(),
n_survived = sum(Survived == 1)) %>% mutate(survived_rate = n_survived/n_passengers)

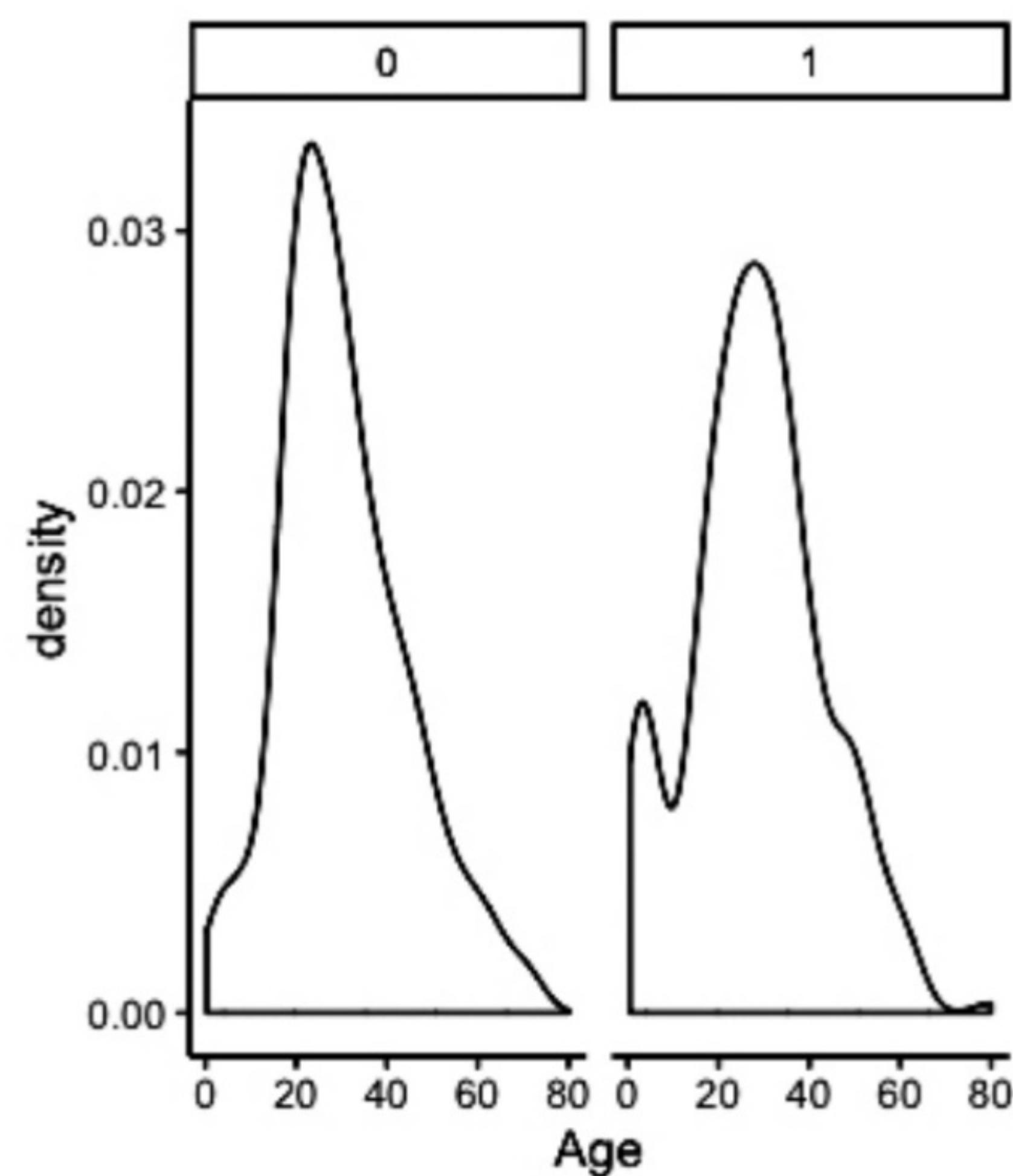
ggplot(stat,aes(x=Sex,y=survived_rate)) +
geom_bar(stat='identity') + theme_bw() + ylab('Survived Rate (%)')
```



[그림 8-31 성별 생존율(%) Plot]

### c) 생존율과 각 변수의 관계 파악 2(연령에 따른 생존율)

```
# Histogram (Survived)
ggplot(data,aes(x=Age)) + geom_density() +
facet_grid(~Survived) + theme_classic()
```

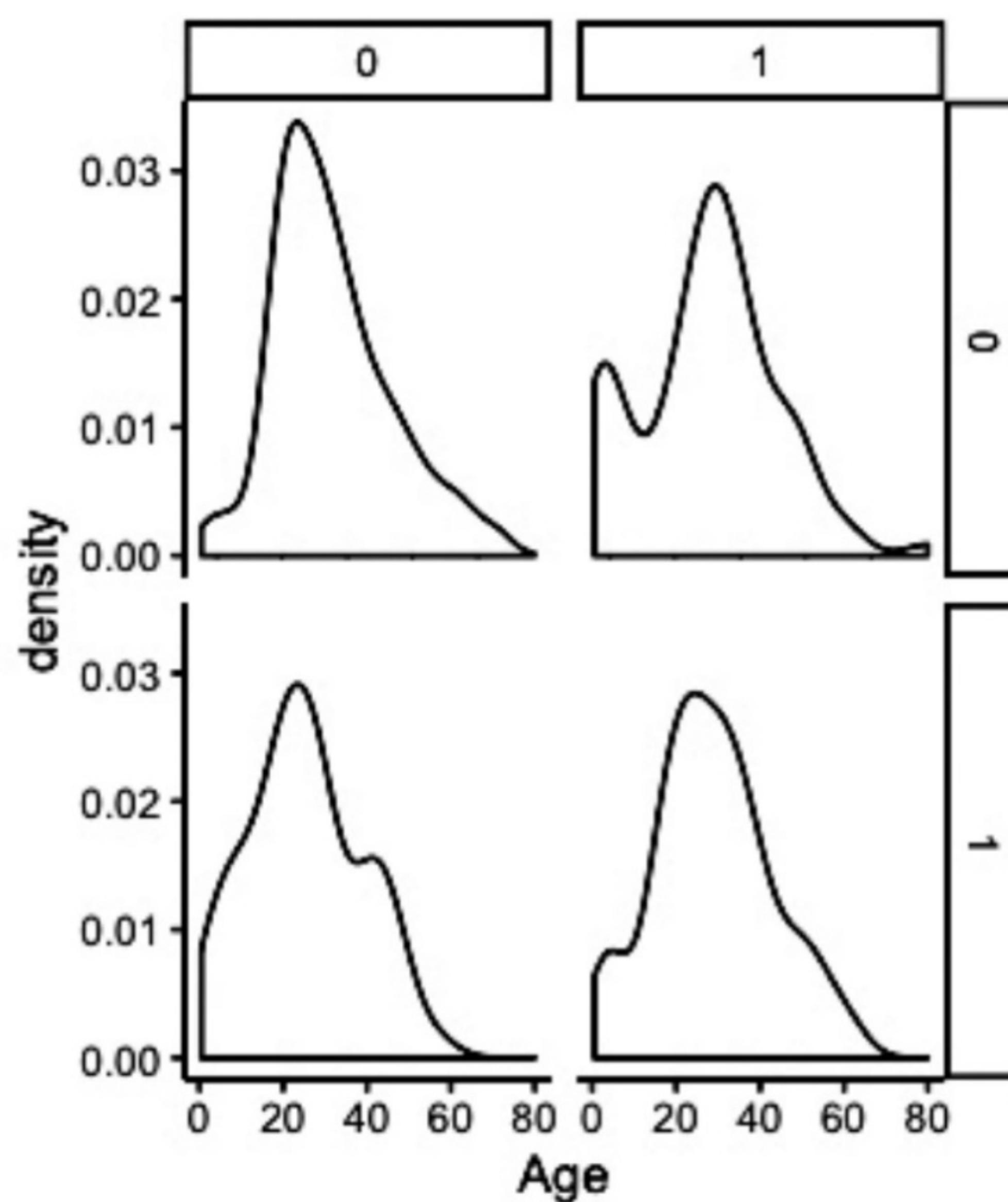


[그림 8-32 연령별 생존율(%) Histogram]

생존자(Survived: 1)와 미 생존자의 연령별 분포를 비교해 보면, 생존자의 경우 유아의 비중이 미 생존자보다 높게 나타나고, 전반적으로 연령대가 높은 것으로 확인된다.

#### d) 생존율과 각 변수의 관계 파악 3(연령, 성별에 따른 생존율)

```
# Histogram (Sex~Survived)
ggplot(data,aes(x=Age)) + geom_density() +
facet_grid(Sex~Survived) + theme_classic()
```



[그림 8-33 연령별, 성별 생존율(%) Histogram]

#### e) 모델 생성

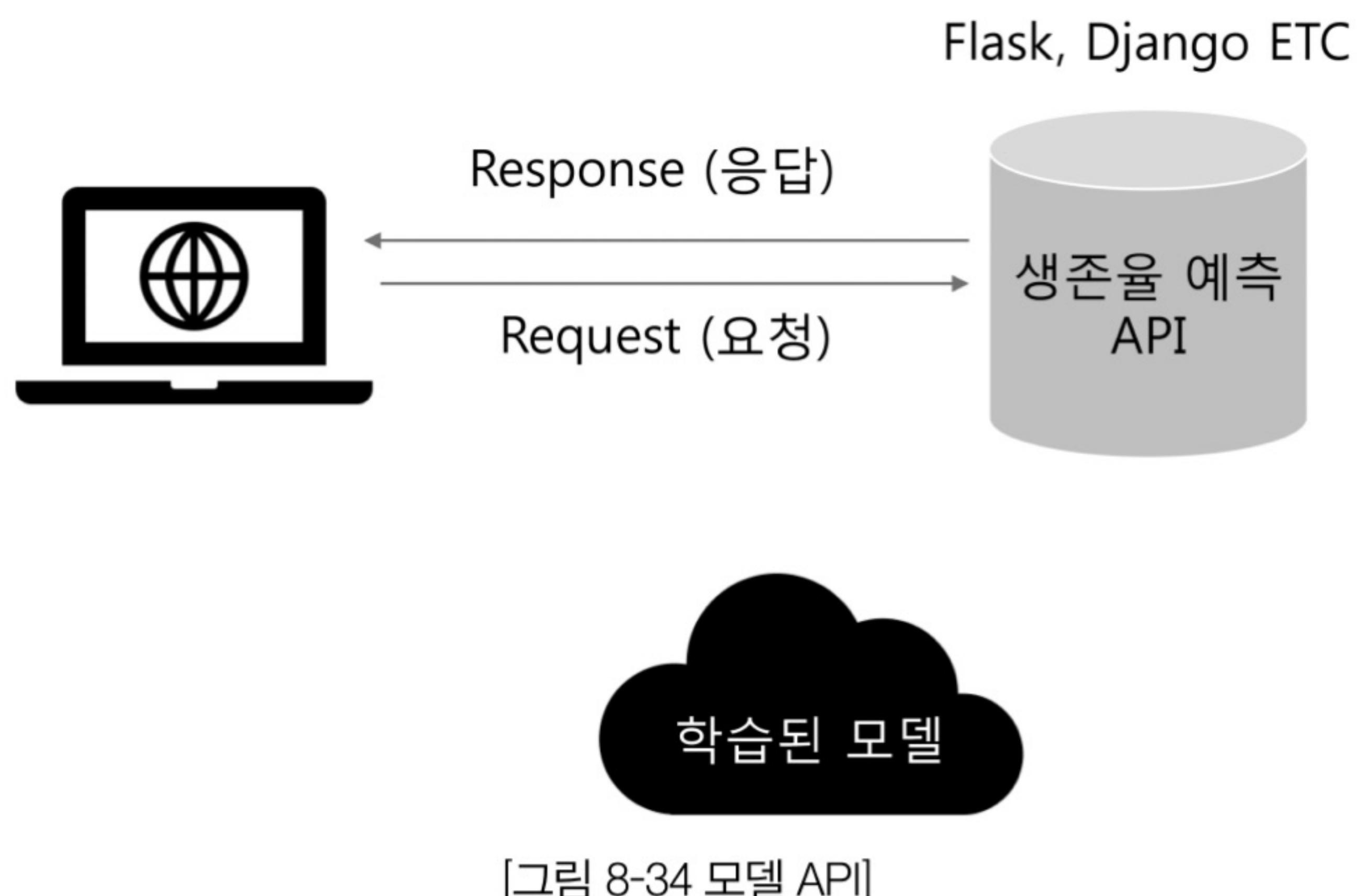
생존 여부를 예측하는 Binary Classification을 이용한다.

#### f) 모델 평가

AUC, ACC, F1 Score 등과 같은 지표를 이용해 모델을 평가한다.

### g) 모델 배포

우리가 생성한 모델을 API로 제공한다면, HTTP 통신을 통해 예측 결과를 가져올 수 있다.



[그림 8-34 모델 API]

Python의 경우 보편적으로 Flask, Django와 같은 프레임워크를 이용해 request에 응답하는 API를 만든다. 생존율 예측 API는 Request의 정보(데이터)를 학습된 모델을 이용해 예측하고, 이를 Response로 반환하게 된다.

머신러닝 모델 생성 과정을 전반적으로 살펴보았다. 사실 더 복잡한 내용이 있지만, 전체적인 흐름을 이해하고, 우리가 학습하는 SQL이 어떻게 사용되는지 이해하면 좋을 것 같다. SQL을 학습하고, 무엇을 공부하는 게 좋을지 고민하는 독자가 있다면, 본인의 관심(데이터 분석, 모델 생성 등)에 맞는 방향으로 효율적으로 학습 하길 바란다.