



데이터 분석



1. 데이터 포맷의 이해



1. 데이터 포맷의 이해

데이터는 그 구조와 형식에 따라 크게 정형, 비정형, 반정형으로 분류할 수 있습니다. 각 데이터 유형은 특정한 특성과 용도를 가지고 있으며, 데이터 분석과 처리 방식도 이에 따라 다르다.

1. 정형 데이터 (Structured Data)

정형 데이터는 고정된 필드와 레코드로 구성된 데이터입니다. 이러한 데이터는 일반적으로 테이블 형식으로 저장되며, 각 필드에는 명확하게 정의된 데이터 유형이 있다.

특징:

- **고정된 구조:** 행과 열로 이루어진 표 형태로 데이터가 정렬된다.
- **명확한 데이터 타입:** 각 열의 데이터 타입이 명확히 정의된다.
- **관계형 데이터베이스에서 사용:** SQL을 통해 데이터베이스에서 관리된다.

예시:

- **스프레드시트:** Excel 파일 (.xlsx)
- **관계형 데이터베이스:** MySQL, PostgreSQL

1. 데이터 포맷의 이해

2. 비정형 데이터 (Unstructured Data)

비정형 데이터는 고정된 구조가 없는 데이터입니다. 텍스트, 이미지, 비디오, 오디오 등 다양한 형식이 포함되며, 분석을 위해 추가적인 처리 및 변환이 필요합니다.

특징:

- **고정된 구조 없음:** 데이터를 저장하는 형식이 일정하지 않음.
- **다양한 형식:** 텍스트 파일, 이미지, 비디오, 오디오 등 다양한 미디어 형식 포함.
- **처리 복잡성:** 분석을 위해서는 자연어 처리(NLP), 이미지 처리 등 고급 기술이 필요.

예시:

- **텍스트 파일:** 로그 파일, 소셜 미디어 포스트
- **멀티미디어 파일:** 이미지 파일 (.jpg, .png), 비디오 파일 (.mp4)

1. 데이터 포맷의 이해

3. 반정형 데이터 (Semi-Structured Data)

반정형 데이터는 정형 데이터와 비정형 데이터의 중간 형태로, 일정한 구조를 가지고 있지만 고정된 스키마는 없는 데이터

특징:

- **유연한 구조:** 데이터의 구조가 존재하지만 고정된 스키마는 없음.
- **계층적 구조:** 중첩된 데이터 구조를 표현할 수 있음.
- **일관성 유지:** 데이터 항목 간의 일관성을 유지하면서도 유연한 확장이 가능.

예시:

- XML (Extensible Markup Language)
- JSON (JavaScript Object Notation)

2. CSV (Comma-Separated Values) 파일

2. CSV 파일

CSV 파일의 특징

CSV는 **Comma-Separated Values**의 약자로, 각 데이터가 쉼표(,)로 구분되는 텍스트 파일 형식이다. CSV 파일은 보통 아래와 같은 형태를 가진다.

id,name,age,city
10,John,28,New York
11,Jane,32,Los Angeles
12,Doe,22,Chicago



id	name	age	city
10	John	28	New York
11	Jane	32	Los Angeles
12	Doe	22	Chicago

여기서 각 줄은 하나의 레코드를 나타내며, 쉼표로 구분된 각 항목은 필드를 나타냅니다. 첫 번째 줄은 일반적으로 열 제목(header)을 포함하여 각 열의 의미를 명확히 합니다.

장점

- 1. 광범위한 지원:** CSV 포맷은 거의 모든 데이터 분석 도구와 프로그래밍 언어에서 지원됩니다. Python, R, Excel, SQL, 그리고 대부분의 데이터베이스 시스템에서 CSV 파일을 쉽게 읽고 쓸 수 있다.
- 2. 사용 용이성:** CSV 파일은 사람이 쉽게 읽고 이해할 수 있는 형식이다. 텍스트 에디터로 열어볼 수 있으며, 복잡한 소프트웨어 없이도 데이터를 검토할 수 있다.
- 3. 가벼운 파일 크기:** 다른 데이터 포맷에 비해 비교적 단순한 구조로 인해 파일 크기가 작다.

2. CSV 파일

단점

- 1.대용량 데이터 처리의 비효율성:** CSV 파일은 구조가 단순하고 인덱스나 메타데이터를 포함하지 않기 때문에 대용량 데이터를 처리할 때 비효율적일 수 있다. 모든 데이터를 메모리에 로드해야 하기 때문에 성능이 저하될 수 있다.
- 2.메타데이터 포함 불가:** CSV 포맷은 데이터 외에 추가적인 메타데이터를 포함할 수 없다. 예를 들어, 데이터 유형, 인덱스, 주석 등의 정보를 저장할 수 없다.
- 3.데이터 정렬의 어려움:** CSV 파일은 계층적 구조를 지원하지 않기 때문에 복잡한 데이터 관계를 표현하기 어렵다. 예를 들어, 중첩된 데이터 구조를 표현하는 데 한계가 있다.
- 4.데이터 무결성 문제:** 쉼표가 데이터 값의 일부로 포함될 경우 데이터의 무결성이 깨질 수 있다. 이를 방지하기 위해 따옴표로 데이터를 감싸거나 이스케이프 문자를 사용하는 등의 추가적인 처리 작업이 필요하다.

요약

CSV는 각 데이터 항목이 쉼표로 구분되는 텍스트 파일 형식이다. 광범위한 지원과 사용 용이성, 가벼운 파일 크기 등의 장점이 있지만, 대용량 데이터 처리 비효율성, 메타데이터 포함 불가 등의 단점도 있다. CSV는 단순한 데이터 전송과 저장에 유용하다.

3. Excel (XLSX) 파일

3. Excel (XLSX) 파일

특징

Excel 파일 형식(XLSX)은 마이크로소프트 엑셀에서 사용하는 스프레드시트 파일 형식. 데이터가 행과 열로 구성된 표 형태로 저장돼 있어. XLSX 파일은 복잡한 계산, 그래프, 피벗 테이블 등의 기능을 지원하면서 데이터를 관리할 수 있는 강력한 도구

장점

- 1.데이터 시각적 편집 및 관리:** 엑셀은 데이터를 시각적으로 편집하고 관리하는 데 매우 편리하며 셀을 클릭하여 데이터를 입력하거나 수정할 수 있고, 다양한 서식 도구를 사용해 데이터를 강조하거나 체계적으로 정리할 수 있다
- 1.여러 시트 포함 가능:** 하나의 XLSX 파일 안에 여러 개의 시트를 포함할 수 있고 이를 통해 서로 관련된 데이터를 하나의 파일에서 분리하고 체계적으로 관리할 수 있다
- 2.고급 기능 지원:** 엑셀은 수식, 매크로, 그래프, 피벗 테이블 등 다양한 고급 기능을 제공하며 이를 통해 데이터 분석과 보고서를 쉽게 작성할 수 있다.
- 3.사용자 친화적 인터페이스:** 엑셀은 직관적이고 사용자 친화적인 인터페이스를 제공하여 데이터 분석 경험이 적은 사람도 쉽게 사용할 수 있다.

3. Excel (XLSX) 파일

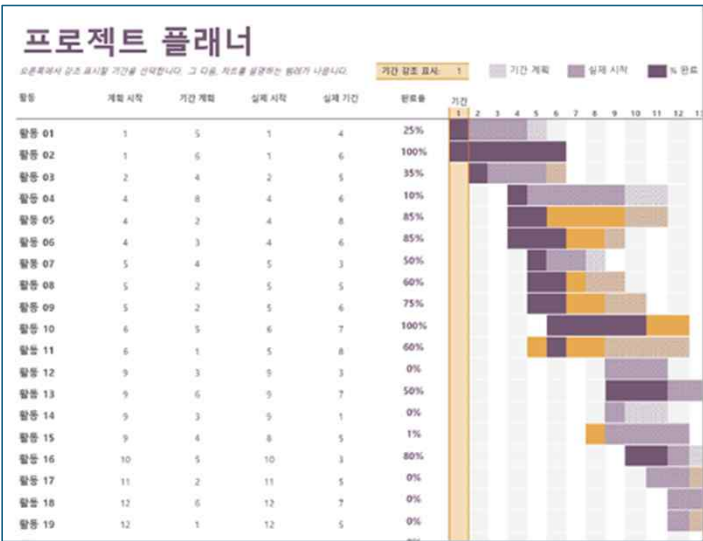
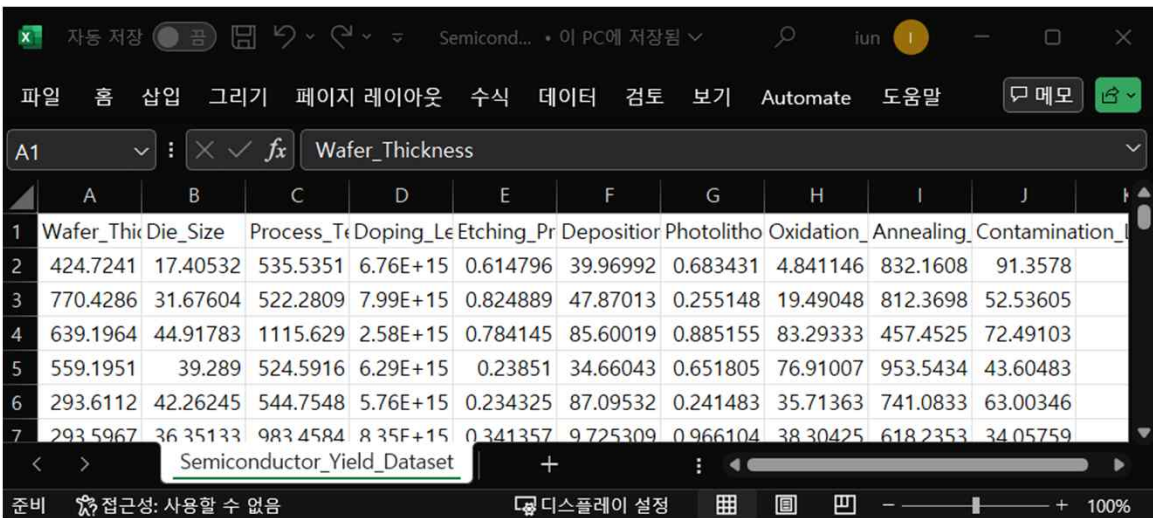
단점

- 1.대용량 데이터 처리 비효율:** 엑셀은 대용량 데이터를 처리할 때 성능이 저하될 수 있고 행과 열의 수가 많아지면 파일 크기가 커지고, 계산 속도가 느려질 수 있다. 대규모 데이터를 다룰 때는 데이터베이스나 빅데이터 도구를 사용하는 것이 더 효율적임.
- 2.호환성 문제:** 엑셀 파일은 다른 시스템이나 소프트웨어와의 호환성이 CSV에 비해 낮을 수 있다. 예를 들어, 특정 소프트웨어나 프로그래밍 언어에서 XLSX 파일을 직접 읽거나 쓰기 어려운 경우가 있다. 이럴 때는 CSV 같은 단순한 텍스트 형식이 더 유리할 수 있다.
- 3.데이터 무결성 위험:** 엑셀 파일은 수동으로 데이터를 입력하고 수정할 수 있다, 실수로 인해 데이터가 손상되거나 무결성이 깨질 위험이 있고 특히 여러 사람이 동일한 파일을 동시에 편집할 때 이런 문제가 발생할 가능성이 높다.
- 4.보안 취약성:** XLSX 파일은 암호를 설정할 수 있지만, 여전히 보안에 취약할 수 있다. 중요한 데이터는 데이터베이스와 같은 더 안전한 저장소에 보관하는 것이 좋을 수 있다.

3. Excel (XLSX) 파일

요약

엑셀 파일(XLSX)은 데이터를 시각적으로 편집하고 관리하기 좋은 스프레드시트 형식이며 여러 시트를 포함할 수 있고 다양한 고급 기능을 제공하지만, 대용량 데이터 처리에 비효율적이고 호환성 문제가 있을 수 있다. 따라서 엑셀은 소규모 데이터 분석 및 보고서 작성에 적합하며, 대규모 데이터 처리를 위해서는 다른 도구를 고려해야 한다.



4. JSON (JavaScript Object Notation) 포맷

4. JSON (JavaScript Object Notation) 포맷

특징

JSON은 JavaScript Object Notation의 약자로, 속성-값 쌍으로 이루어진 텍스트 형식.

이 형식은 사람이 읽고 쓰기 쉽고, 기계가 해석하고 생성하기도 용이한 데이터 교환 형식이다. JSON 데이터는 중괄호 {}와 대괄호 []를 사용해 계층적 구조를 표현할 수 있다

```
{  
  "name": "John Doe",  
  "age": 30,  
  "is_student": false,  
  "courses": ["Math", "Science", "Literature"],  
  "address": {  
    "street": "123 Main St",  
    "city": "Anytown",  
    "zipcode": "12345"  
  }  
}
```

JSON 데이터는 name, age, is_student, courses, address 등의 속성-값 쌍으로 구성돼 있어. 각 속성은 콜론(:)을 사용해 값과 연결되며, 값은 문자열, 숫자, 배열, 객체 등 다양한 데이터 유형이 될 수 있다

4. JSON (JavaScript Object Notation) 포맷

장점

- 1. 구조화된 데이터 표현에 적합:** JSON은 데이터를 계층적이고 구조적으로 표현하는 데 매우 적합하며 중첩된 객체와 배열을 통해 복잡한 데이터 구조를 쉽게 나타낼 수 있다.
- 2. 웹 API와의 호환성:** JSON은 웹 API에서 데이터 교환 형식으로 널리 사용된다. 대부분의 웹 서비스는 JSON 형식을 사용해 데이터를 송수신하며, 이는 JSON이 JavaScript와의 호환성이 뛰어나기 때문이다.
- 3. 언어 독립성:** JSON은 JavaScript 기반으로 설계됐지만, 대부분의 프로그래밍 언어에서 지원되며. Python, Java, C#, PHP 등 다양한 언어에서 JSON 데이터를 쉽게 파싱하고 생성할 수 있다.
- 4. 가벼운 데이터 포맷:** JSON은 텍스트 기반 데이터 포맷으로, 비교적 가벼운 크기를 가지고 있어 네트워크를 통해 데이터를 전송할 때 효율적이다.

4. JSON (JavaScript Object Notation) 포맷

단점

- 1. 대용량 데이터 처리 속도 저하:** JSON은 대용량 데이터의 경우 처리 속도가 느려질 수 있다. JSON 파싱은 메모리 소모가 크기 때문에, 매우 큰 JSON 파일을 처리할 때 성능이 저하될 수 있다. 대규모 데이터를 다룰 때는 바이너리 포맷을 고려할 수 있다.
- 2. 스키마가 없음:** JSON은 정해진 스키마가 없어 데이터의 구조가 유연하지만, 이는 데이터 무결성을 보장하는 데 어려움을 줄 수 있다. 데이터의 일관성을 유지하려면 추가적인 검증 작업이 필요하다.
- 3. 이진 데이터 비효율:** JSON은 텍스트 기반 포맷이기 때문에, 이진 데이터를 저장할 때 비효율적일 수 있다. 이진 데이터를 처리할 때는 Base64 인코딩을 사용해야 하며, 이는 데이터 크기를 증가시키고 처리 속도를 늦출 수 있다.

4. JSON (JavaScript Object Notation) 포맷

요약

JSON은 속성-값 쌍으로 이루어진 텍스트 형식의 데이터 포맷이다. 구조화된 데이터를 표현하는 데 적합하고, 웹 API와의 호환성이 뛰어나며, 다양한 프로그래밍 언어에서 지원된다. 그러나 대용량 데이터를 처리할 때 성능이 저하될 수 있고,スキーマ가 없어 데이터 일관성을 유지하기 어렵다. JSON은 가벼운 데이터 교환 포맷으로, 특히 웹 개발에서 널리 사용된다.

5. SQL (Structured Query Language) 포맷

5. SQL (Structured Query Language) 포맷

특징

SQL은 **Structured Query Language**의 약자로, 관계형 데이터베이스에서 데이터를 관리하고 조작하기 위해 사용되는 질의 언어이다. SQL을 통해 데이터베이스에서 데이터를 검색, 삽입, 업데이트, 삭제할 수 있으며, 데이터베이스 스키마를 정의하고 수정할 수도 있다.

데이터베이스 테이블 생성

```
sql

CREATE TABLE students (
  id INT PRIMARY KEY,
  name VARCHAR(50),
  age INT,
  major VARCHAR(50)
);
```

데이터 삽입

sql

코드 복사

```
INSERT INTO students (id, name, age, major) VALUES (1, 'John Doe', 20, 'Computer Science');
INSERT INTO students (id, name, age, major) VALUES (2, 'Jane Smith', 22, 'Mathematics');
INSERT INTO students (id, name, age, major) VALUES (3, 'Michael Brown', 21, 'Physics');
INSERT INTO students (id, name, age, major) VALUES (4, 'Emily Davis', 23, 'Chemistry');
INSERT INTO students (id, name, age, major) VALUES (5, 'Daniel Wilson', 20, 'Biology');
INSERT INTO students (id, name, age, major) VALUES (6, 'Sophia Johnson', 22, 'Economics');
INSERT INTO students (id, name, age, major) VALUES (7, 'David Lee', 21, 'History');
INSERT INTO students (id, name, age, major) VALUES (8, 'Emma Garcia', 23, 'Literature');
INSERT INTO students (id, name, age, major) VALUES (9, 'Olivia Martinez', 20, 'Art');
INSERT INTO students (id, name, age, major) VALUES (10, 'James Anderson', 22, 'Philosophy');
```

5. SQL (Structured Query Language) 포맷

데이터 조회

sql

```
SELECT * FROM students;
```

조회 결과

id	name	age	major
1	John Doe	20	Computer Science
2	Jane Smith	22	Mathematics
3	Michael Brown	21	Physics
4	Emily Davis	23	Chemistry
5	Daniel Wilson	20	Biology
6	Sophia Johnson	22	Economics
7	David Lee	21	History
8	Emma Garcia	23	Literature
9	Olivia Martinez	20	Art
10	James Anderson	22	Philosophy

5. SQL (Structured Query Language) 포맷

장점

- 1. 대규모 데이터 관리에 최적화:** SQL은 대규모 데이터를 효율적으로 관리하는 데 최적화되어 있다. 관계형 데이터베이스는 인덱싱, 조인, 트랜잭션 관리 등의 기능을 제공하여 대용량 데이터를 신속하고 정확하게 처리할 수 있다.
- 2. 강력한 쿼리 기능:** SQL은 매우 강력한 쿼리 기능을 제공한다. SELECT, INSERT, UPDATE, DELETE 등의 기본 명령어를 통해 데이터를 조작할 수 있으며, JOIN, SUBQUERY, AGGREGATE FUNCTIONS 등을 사용하여 복잡한 쿼리도 쉽게 작성할 수 있다.
- 3. 데이터 무결성과 보안:** SQL은 데이터 무결성과 보안을 보장하는 다양한 메커니즘을 제공한다. PRIMARY KEY, FOREIGN KEY, UNIQUE 등의 제약 조건을 통해 데이터의 일관성을 유지할 수 있으며, 권한 관리와 트랜잭션 제어를 통해 데이터베이스의 보안을 강화할 수 있다.
- 4. 표준화된 언어:** SQL은 국제 표준화 기구(ISO)와 미국 표준화 기구(ANSI)에 의해 표준화되어 있다. 이는 SQL이 다양한 데이터베이스 시스템에서 일관되게 사용될 수 있음을 의미하며, 사용자가 특정 데이터베이스 시스템에 종속되지 않고도 SQL을 배워서 활용할 수 있게 한다.

5. SQL (Structured Query Language) 포맷

단점

1.비구조화된 데이터 처리에 부적합: SQL은 관계형 데이터베이스의 테이블 구조에 최적화되어 있기 때문에, 비구조화된 데이터나 반정형 데이터를 처리하는 데 부적합하다. 텍스트 파일, JSON, XML과 같은 데이터 형식은 SQL로 직접 처리하기 어렵다. 이러한 데이터는 NoSQL 데이터베이스나 다른 데이터 처리 도구를 사용하는 것이 더 적합하다.

2.복잡한 스키마 변경의 어려움: SQL 데이터베이스는 스키마가 고정되어 있어, 스키마를 변경하는 작업이 복잡하고 시간이 많이 소요될 수 있다. 특히 대규모 데이터베이스에서 스키마를 변경하려면 데이터베이스를 중지하고 변경 작업을 수행해야 하는 경우도 있어, 운영에 큰 영향을 줄 수 있다.

3.확장성의 한계: SQL 데이터베이스는 수직적 확장(성능 향상을 위해 더 강력한 서버를 사용하는 것)에 최적화되어 있지만, 수평적 확장(여러 서버로 데이터베이스를 분산하는 것)에는 한계가 있다. 반면, NoSQL 데이터베이스는 수평적 확장에 더 적합하다.

4.복잡한 쿼리의 성능 저하: 매우 복잡한 쿼리나 다중 테이블 조인을 사용할 경우, 성능이 저하될 수 있다. 이러한 문제를 해결하기 위해 인덱스 튜닝, 쿼리 최적화 등의 추가 작업이 필요하다.

5. SQL (Structured Query Language) 포맷

요약

SQL은 관계형 데이터베이스에서 데이터를 관리하고 조작하는 데 사용되는 강력한 질의 언어이다. 대규모 데이터 관리와 쿼리에 최적화되어 있으며, 데이터 무결성과 보안을 보장하는 다양한 기능을 제공한다. 하지만 비구조화 된 데이터 처리에 부적합하고, 스키마 변경과 확장성에 한계가 있을 수 있다. SQL은 표준화된 언어로서 다양한 데이터베이스 시스템에서 일관되게 사용할 수 있다.

6. XML (Extensible Markup Language) 포맷

6. XML (Extensible Markup Language) 포맷

특징

XML은 **Extensible Markup Language**의 약자로, 데이터를 구조화하기 위해 태그를 사용하여 정의된 텍스트 기반 데이터 형식이다. XML은 사람이 읽고 쓰기 쉽도록 설계되었으며, 다양한 시스템 간에 데이터를 교환하는 데 사용된다. XML은 자체적으로 태그를 정의할 수 있어 유연하고 확장 가능하다.

```
<students>
  <student>
    <id>1</id>
    <name>John Doe</name>
    <age>20</age>
    <major>Computer Science</major>
  </student>
  <student>
    <id>2</id>
    <name>Jane Smith</name>
    <age>22</age>
    <major>Mathematics</major>
  </student>
</students>
```

이 예제는 students라는 루트 엘리먼트 아래 여러 student 엘리먼트를 포함하고 있다. 각 student 엘리먼트는 id, name, age, major 태그를 통해 학생의 정보를 표현하고 있다.

6. XML (Extensible Markup Language) 포맷

장점

- 1. 유연한 데이터 구조:** XML은 사용자가 직접 태그를 정의할 수 있어, 다양한 형태의 데이터를 표현하는 데 유연하다. 계층적 구조를 쉽게 표현할 수 있어 복잡한 데이터를 잘 나타낼 수 있다.
- 2. 시스템 간 호환성:** XML은 플랫폼과 언어에 독립적이어서, 다양한 시스템 간에 데이터를 교환하는 데 적합하다. 웹 서비스와 데이터 교환 포맷으로 널리 사용된다.
- 3. 표준화된 형식:** XML은 W3C(World Wide Web Consortium)에 의해 표준화되어 있어, 일관된 데이터 표현 방식을 제공한다. 이를 통해 데이터를 교환하는 시스템 간의 호환성을 보장한다.
- 4. 데이터 검증 가능:** XML은 DTD(Document Type Definition)나 XML Schema를 통해 문서의 구조를 정의하고 검증할 수 있다. 이를 통해 데이터의 일관성과 무결성을 유지할 수 있다.

6. XML (Extensible Markup Language) 포맷

단점

- 1. 대용량 데이터의 비효율성:** XML은 태그를 사용하여 데이터를 표현하기 때문에, 데이터 자체보다 태그가 차지하는 공간이 더 클 수 있다. 이는 대용량 데이터를 다룰 때 비효율적일 수 있다.
- 2. 복잡한 파싱:** XML 데이터를 파싱하는 작업은 비교적 복잡하고 리소스를 많이 소모할 수 있다. 특히 큰 XML 문서를 처리할 때 성능이 저하될 수 있다.
- 3. 가독성의 한계:** 사람이 읽을 수 있는 형식으로 설계되었지만, 복잡한 구조의 XML 문서는 가독성이 떨어질 수 있다. 태그가 많아지면 문서가 길어지고 읽기 어려워질 수 있다.
- 4. 구조적 제약:** XML은 엄격한 구조적 규칙을 따르기 때문에, 유연성이 떨어질 수 있다. 비정형 데이터나 매우 동적인 데이터 구조를 표현하는 데는 적합하지 않을 수 있다.

6. XML (Extensible Markup Language) 포맷

요약

XML은 데이터를 구조화하기 위해 태그를 사용하여 정의된 텍스트 기반 데이터 형식이다. 유연한 데이터 구조와 시스템 간 호환성, 표준화된 형식, 데이터 검증 가능 등의 장점을 가지고 있지만, 대용량 데이터의 비효율성, 복잡한 파싱, 가독성의 한계, 구조적 제약 등의 단점도 있다. XML은 다양한 시스템 간에 데이터를 교환하는데 유용한 포맷이다.

7. 데이터 로딩 실습



Gemini로 CSV파일 업로드하기

- 데이터셋 로드 및 확인: "boston_train.csv" CSV 파일을 불러오고 데이터셋 정보를 확인.

boston_train

CSV

데이터 설명해줘

코드 표시 <>

◆ 분석 ▼

제공해주신 파일은 **보스턴 주택 가격 예측 데이터셋(Boston Housing Dataset)**으로 보입니다.

이 데이터셋은 총 **400개의 행**과 **10개의 열**로 구성되어 있으며, 모든 열에 결측값 없이 채워져 있습니다. 대부분의 열은 ****실수형(float64)****이며, 'TAX' 열만 ****정수형(int64)****입니다.

주요 열에 대한 설명은 다음과 같습니다.

보스턴 주택 가격 데이터셋 주요 변수

열 이름	설명	단위 및 참고 사항
CRIM	도시별 1인당 범죄율	
ZN	25,000 ft ² 초과 주거 지역의 비율	
INDUS	도시별 비소매 상업 지구의 비율	
NOX	일산화질소 농도	1천만 분율 (parts per 10 million)
RM	주택당 평균 방 개수	
AGE	1940년 이전에 건축된 주택의 비율	
DIS	보스턴 5개 고용 센터까지의 가중 거리	

자세히 분석해줘

코드 표시 <>

◆ 분석 ▼

보스턴 주택 가격 데이터셋(boston_train.csv)에 대한 상세 분석 결과입니다.

분석은 **기술 통계(Descriptive Statistics)**를 통한 데이터 분포 파악과 **목표 변수(MEDV)와의 상관관계 분석**을 중심으로 진행되었습니다.

1. 기술 통계 분석 (Descriptive Statistics)

각 변수의 평균, 표준편차, 최솟값, 최댓값 등을 통해 데이터의 전반적인 규모와 분포를 파악할 수 있습니다.

변수	평균 (Mean)	표준편차 (Std)	최솟값 (Min)	25% (Q1)	중앙값 (50%, Median)	최댓값 (Max)
CRIM (범죄율)	3.91	9.37	0.01	0.08	0.28	88.98
ZN (주거지 비율)	11.35	23.51	0.00	0.00	0.00	100.00
RM (평균 방 개수)	6.28	0.73	3.56	5.88	6.21	8.78
AGE (오래된 주택 비율)	69.11	28.52	2.90	44.85	78.80	100.00
TAX (재산세율)	411.65	171.63	187.00	277.00	332.00	711.00
PTRATIO (학생-교사 비율)	18.49	2.16	12.60	17.40	19.10	22.00
MEDV (주택 가격 중앙값)	22.63	9.57	5.00	16.60	21.40	50.00

엑셀(xlsx)파일 업로드하기 : credit_card.xlsx

credit_card

XLSX

데이터 설명해줘(한글로)

코드 표시 <>

◆ 분석 ▼

제공해주신 파일은 **대만 신용카드 고객의 연체 예측 데이터셋**으로 보입니다.

총 **30,000개의 행**과 **24개의 열**로 구성되어 있으며, 모든 열은 결측값 없이 **정수형(int64)** 데이터로 이루어져 있습니다.

주요 열에 대한 설명은 다음과 같습니다.

📄 고객 정보 및 신용 한도 (Customer Profile and Limit)

열 이름	설명
LIMIT_BAL	신용카드 발급액 (Credit limit)
SEX	성별 (1=남자, 2=여자)
EDUCATION	학력 (1=대학원, 2=대학교, 3=고등학교, 4=기타, 5, 6=알 수 없음)
MARRIAGE	결혼 상태 (1=기혼, 2=미혼, 3=기타)
AGE	나이

과거 상환 상태 (Historical Repayment Status)

- PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6: 각각 2005년 9월부터 4월까지의 월별 상환 상태를 나타냅니다.
 - 0: 기한 내 납부 (Paid on time)
 - 1: 1개월 연체 (Delayed 1 month)
 - 2: 2개월 연체 (Delayed 2 months)
 - ...
 - -1: 전월 전액 납부 (Paid in full)
 - -2: 해당 월에 거래 없음 (No consumption)

청구 금액 및 지불 금액 (Bill and Payment Amounts)

열 이름	설명
BILL_AMT1 ~ BILL_AMT6	2005년 9월부터 4월까지의 월별 청구 금액 (청구서 금액)
PAY_AMT1 ~ PAY_AMT6	2005년 9월부터 4월까지의 월별 지불 금액 (이전 달 청구서에 대한 지불 금액)

예측 대상 (Target Variable)

열 이름	설명
default payment next month	다음 달 (2005년 10월) 신용카드 대금 연체 여부입니다. 예측하고자 하는 목표 변수입니다. (1=연체, 0=연체 안 함)

8. 탐색적 데이터 분석(EDA)

Exploratory Data Analysis



WHO 국가 복지 데이터셋 사용

• 데이터셋 로드 및 확인 :
WHO_first9cols.csv CSV파일을 불러오고
기본 정보를 확인.

프롬프트 : "이 데이터셋에 대한 정보를
간단히 알려줘"

분석 ▾

제공해주신 파일은 WHO (세계보건기구) 데이터를 기반으로 한 국가별 사회 및 경제 지표 데이터셋으로 보입니다.

데이터 요약

- 크기: 총 202개 국가에 대한 데이터가 포함되어 있습니다.
- 열 개수: 총 9개의 열이 있습니다.
- 특징: 국가명, 대륙 정보, 청소년 출산율, 성인 문해율, 국민 총소득, 초등학교 순 등록률, 총인구 등의 지표를 포함하고 있습니다.
- 특이 사항: 여러 변수에서 결측값(NaN)이 존재합니다. 예를 들어, 성인 문해율(Adult literacy rate)은 전체 202개 국가 중 131개 국가의 데이터만 있습니다.

주요 열 설명

열 이름 (번역)	원래 영어 이름	데이터 유형
국가	Country	문자열 (Object)
대륙	Continent	정수형 (int64)
청소년 출산율 (%)	Adolescent fertility rate (%)	실수형 (float64)
성인 문해율 (%)	Adult literacy rate (%)	실수형 (float64)
1인당 국민 총소득 (PPP 국제 \$)	Gross national income per capita (PPP international \$)	실수형 (float64)
초등학교 순 등록률 (여)	Net primary school enrolment ratio female (%)	실수형 (float64)
초등학교 순 등록률 (남)	Net primary school enrolment ratio male (%)	실수형 (float64)
총인구 (천 명)	Population (in thousands) total	실수형 (float64)

기초 통계 값을 알려줘.

◆ 분석 ▼

WHO 국가별 지표 데이터셋의 기초 통계 분석 결과입니다.

분석 결과는 국가 $\text{\texttt{ID}}$ 와 대륙 코드와 같은 분류 변수를 포함하여 모든 수치형 변수에 대해 총 개수, 평균, 표준편차, 최솟값, 4분위수(25%, 50%, 75%), 최댓값을 제공합니다.

📊 주요 통계값 요약 (국가 $\text{\texttt{202}}$ 개 기준)

변수	총 개수 (Count)	평균 (Mean)	표준편차 (Std)	최솟값 (Min)	중앙값 (Median, $\text{\texttt{50\%}}$)	최댓값 (Max)
청소년 출산율 (%)	177	59.46	49.11	0	46	199
성인 문해율 (%)	131	78.87	20.42	23.6	86.5	99.8
1인당 국민 총 소득 (PPP \$)	178	11,250	12,587	260	6,175	60,870
초등학교 순 등록률 (여) (%)	179	84.03	17.79	6	90	100
초등학교 순 등록률 (남) (%)	179	85.70	15.45	11	90	100
총인구 (천 명)	189	34,099	131,838	2	6,640	1,328,470

주요 관찰 사항

1. **결측값:** 성인 문해율(Adult literacy rate)은 $\text{\$}\text{202}\text{\$}$ 개 국가 중 $\text{\$}\text{131}\text{\$}$ 개 국가 ($\approx 65\%$)만 데이터가 있어 결측값 비율이 높습니다.
2. **소득 편차:** 1인당 국민 총소득(Gross national income)의 평균은 $\text{\$}\text{11,250}\text{\$}$ 이지만, 표준편차($\text{\$}\text{12,587}\text{\$}$)가 매우 커서 국가 간 소득 격차가 매우 심하며, 상위 25% 국가의 소득($\text{\$}\text{14,502.5}\text{\$}$)이 중앙값($\text{\$}\text{6,175}\text{\$}$)의 두 배가 넘습니다.
3. **인구 분포:** 총인구(Population)는 평균 $\text{\$}\text{34}\text{\$}$ 백만 명에 비해 최댓값이 약 $\text{\$}\text{13}\text{\$}$ 억 $\text{\$}\text{28}\text{\$}$ 백만 명으로 매우 심한 편향을 보이며, 이는 중국, 인도 등의 인구가 많은 국가가 포함되어 있기 때문입니다. 중앙값은 $\text{\$}\text{6.64}\text{\$}$ 백만 명입니다.
4. **교육 지표:** 초등학교 순 등록률은 남녀 모두 평균이 84% 이상으로 비교적 높으나, 최솟값이 **여성 6%, 남성 11%**인 국가가 존재하여 교육 접근성의 불균형이 나타납니다.

Sheet로 내보내기로 다운로드후 Sheet를 열어본다.

기초 통계값을 알려줘. ☆ 📁 ☁

파일 수정 보기 삽입 서식 데이터 도구 확장 프로그램 도움말 Gemini

🔍 메뉴 ↶ ↷ 🏠 📊 100% ▾ | W % .0 .00 123 | 기본값 ... ▾ | - 10 + | B I ↺ A | 🎨 📐 📏

K36 ▾ | fx

	A	B	C	D	E	F	G	H
1	변수	총 개수 (Count)	평균 (Mean)	표준편차 (Std)	최솟값 (Min)	중앙값 (Median)	최댓값 (Max)	
2	청소년 출산율 (%)	177	59.46	49.11	0	46	199	
3	성인 문해율 (%)	131	78.87	20.42	23.6	86.5	99.8	
4	1인당 국민 총소득 (PPP \$)	178	11,250	12,587	260	6,175	60,870	
5	초등학교 순 등록률 (여) (%)	179	84.03	17.79	6	90	100	
6	초등학교 순 등록률 (남) (%)	179	85.7	15.45	11	90	100	
7	총인구 (천 명)	189	34,099	131,838	2	6,640	1,328,470	
8								
9								

성인 문해율이 가장 낮은 나라는?

◆ 분석 ▼

성인 문해율(Adult literacy rate)이 가장 낮은 나라는 **부르키나파소(Burkina Faso)**입니다.

성인 문해율은 **23.6%**로 확인되었습니다.

👍 🔄 🌐 📄 ⋮

성인 문해율이 가장 높은 나라는?

◆ 분석 ▼

성인 문해율(Adult literacy rate)이 가장 높은 나라는 쿠바(Cuba)와 에스토니아(Estonia) 두 나라입니다.

두 국가 모두 성인 문해율은 **99.8%**로 확인되었습니다. 🇵🇪

👍 🔄 🌐 📄 ⋮

9. 데이터 시각화



주요 시각적 탐색 도구

1. 그래프 (Graphs)

- (1) 산점도 (Scatter Plot): 두 변수 간의 관계를 시각화한다. 예를 들어, 나이와 소득 간의 관계를 나타낼 수 있다.
- (2) 선 그래프 (Line Graph): 시간에 따른 데이터 변화를 시각화한다. 예를 들어, 월별 매출 변화를 보여줄 수 있다.

2. 차트 (Charts)

- (1) 막대그래프 (Bar Chart): 범주형 데이터의 빈도나 비율을 비교한다. 예를 들어, 각 제품의 판매량을 비교할 수 있다.
- (2) 파이 차트 (Pie Chart): 전체에 대한 각 부분의 비율을 보여준다. 예를 들어, 예산의 항목별 비율을 나타낼 수 있다.



3. 플롯 (Plots)

(1) **히스토그램 (Histogram)**: 연속형 데이터의 분포를 시각화한다.

예를 들어, 학생들의 시험 점수 분포를 나타낼 수 있다.


(2) **박스 플롯 (Box Plot)**: 데이터의 분포와 이상치를 시각화한다.

예를 들어, 다양한 지역의 주택 가격 분포를 비교할 수 있다.



Titanic 생존자 데이터 셋 분석

<https://www.kaggle.com/c/titanic>



Titanic - Machine Learning from Disaster


Start here! Predict survival on the Titanic and get familiar with ML basics

Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Overview


∞ This competition runs indefinitely with a rolling leaderboard. [Learn more.](#)

Description

 **Ahoy, welcome to Kaggle! You're in the right place.**

This is the legendary Titanic ML competition – the best, first challenge for you to dive into ML competitions and familiarize yourself with how the Kaggle platform works.

If you want to talk with other users about this competition, come join our [Discord!](#) We've got channels for competitions, job postings and career discussions, resources, and socializing with your fellow data scientists. Follow the link here:

Competition Host
Kaggle 

Prizes & Awards
Knowledge
Does not award Points or Medals

Participation
1,293,570 Entrants
16,184 Participants
15,976 Teams
59,510 Submissions

Tags

Binary Classification
Tabular Beginner

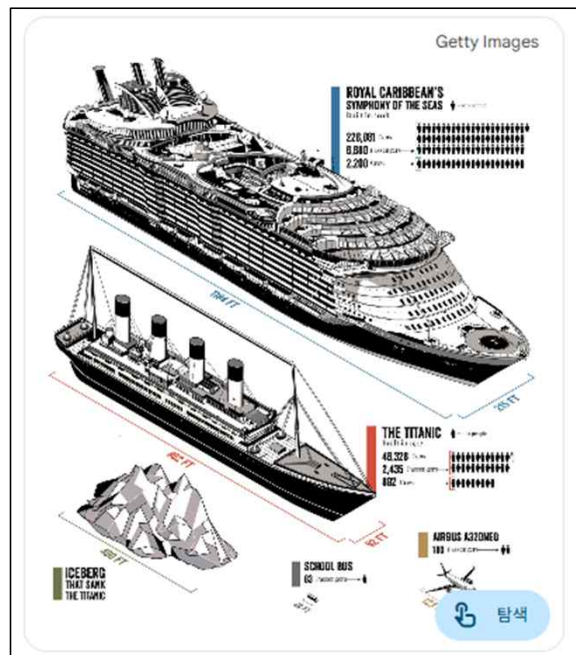
titanic_train.csv 파일 엑셀에서 확인.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mrs	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, Mrs	female	26	0	0	STON/O2. 3101282	7.925		S
5	4	1	1	Futrelle, Mrs	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, Mr	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, Mr	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, Mrs	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, Mrs	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstrom, Mrs	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, Mrs	female	58	0	0	113783	26.55	C103	S
14	13	0	3	Saunders, Mr	male	20	0	0	A/5. 2151	8.05		S
15	14	0	3	Andersson, Mr	male	39	1	5	347082	31.275		S
16	15	0	3	Vestrom, Mrs	female	14	0	0	350406	7.8542		S
17	16	1	2	Hewlett, Mrs	female	55	0	0	248706	16		S
18	17	0	3	Rice, Master	male	2	4	1	382652	29.125		Q
19	18	1	2	Williams, Mr	male		0	0	244373	13		S
20	19	0	3	Vander Planck, Mrs	female	31	1	0	345763	18		S
21	20	1	3	Masella, Mrs	female		0	0	2649	7.225		C
22	21	0	2	Fynney, Mr	male	35	0	0	239865	26		S
23	22	1	2	Beesley, Mr	male	34	0	0	248698	13	D56	S
24	23	1	3	McGowan, Mrs	female	15	0	0	330923	8.0292		Q
25	24	1	1	Sloper, Mr	male	28	0	0	113788	35.5	A6	S
26	25	0	3	Palsson, Mrs	female	8	3	1	349909	21.075		S

Titanic 생존자 데이터 셋 사용

- 데이터셋 로드 및 확인: titanic_train.csv 파일을 불러오고 기본 정보를 확인.

프롬프트 : "이 데이터셋에 대한 정보를 간단히 알려줘."



분석 ▼

제공해주신 파일은 타이타닉호 탑승객의 생존 여부 예측 데이터셋입니다.

이 데이터셋은 타이타닉 침몰 사고 당시 탑승객들의 정보를 담고 있으며, 총 891명의 탑승객과 12개의 변수로 구성되어 있습니다.

데이터 요약

열 이름 (번역)	원래 영어 이름	설명	특이 사항
생존 여부	Survived	생존 (1) 또는 사망 (0)	목표 변수
객실 등급	Pclass	1등석, 2등석, 3등석	범주형
성별	Sex	남성 (male) 또는 여성 (female)	범주형
나이	Age	탑승객의 나이	177개의 결측값
동반 형제/배우자 수	SibSp	동반한 형제 또는 배우자의 수	
동반 부모/자녀 수	Parch	동반한 부모 또는 자녀의 수	
운임	Fare	탑승 요금	
객실 번호	Cabin	객실 번호	결측값이 매우 많음 (204개만 존재)
승선 항구	Embarked	탑승 항구 (C, Q, S)	2개의 결측값

Sheets로 내보내기



이 데이터는 성별, 객실 등급, 나이 등이 생존에 미친 영향을 분석하는 데 주로 사용됩니다.

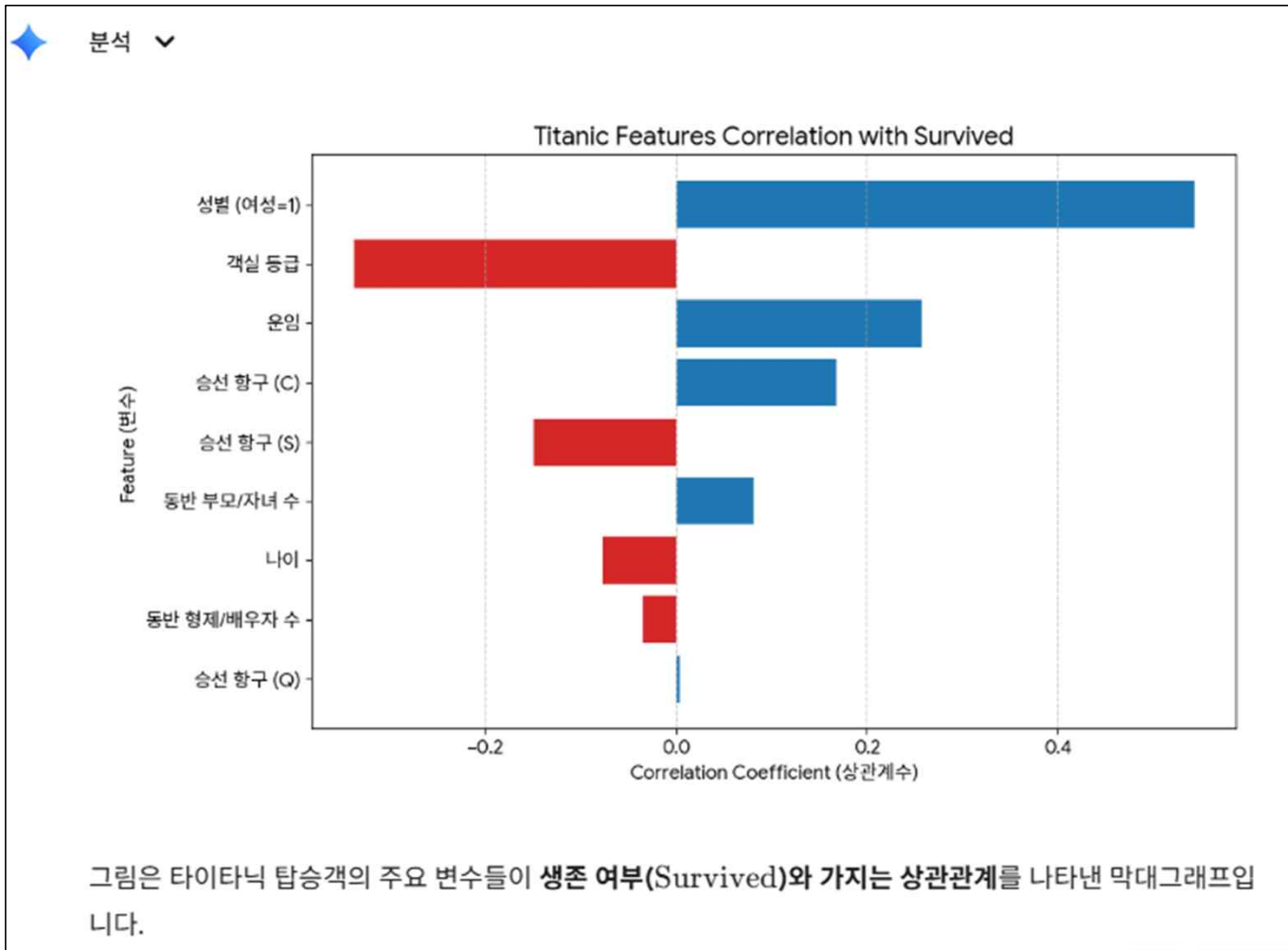
12개 컬럼 모두 설명해줘.

◆ 타이타닉 데이터셋(titanic_train.csv)에 포함된 12개 컬럼에 대한 상세 설명입니다.

번호	컬럼명 (English)	데이터 유형	설명	주요 값 및 특징
1	PassengerId	정수 (int64)	각 탑승객에게 부여된 고유 ID 번호 입니다.	\$\text{1}\$부터 \$\text{891}\$까지의 고유 번호
2	Survived	정수 (int64)	탑승객의 생존 여부 입니다.	0 = 사망, 1 = 생존 (목표 변수)
3	Pclass	정수 (int64)	탑승객이 속한 객실 등급 입니다.	1 = 1등석, 2 = 2등석, 3 = 3등석
4	Name	문자열 (object)	탑승객의 이름 입니다. 이름 안에 호칭(예: Mr., Mrs., Miss) 정보가 포함되어 있습니다.	고유한 문자열 값
5	Sex	문자열 (object)	탑승객의 성별 입니다.	male (남성), female (여성)
6	Age	실수 (float64)	탑승객의 나이 입니다.	\$\text{0.42}\$세부터 \$\text{80}\$세까지의 범위. 177개의 결측값이 존재합니다.
7	SibSp	정수 (int64)	탑승객과 함께 탑승한 형제(Siblings) 또는 배우자(Spouses) 의 수입니다.	\$\text{0}\$부터 \$\text{8}\$까지의 정수
8	Parch	정수 (int64)	탑승객과 함께 탑승한 부모(Parents) 또는 자녀(Children) 의 수입니다.	\$\text{0}\$부터 \$\text{6}\$까지의 정수
9	Ticket	문자열 (object)	탑승객의 티켓 번호 입니다.	다양한 문자열 및 숫자 조합
10	Fare	실수 (float64)	탑승객이 지불한 **운임(요금)** 입니다.	\$\text{0}\$부터 \$\text{512.3292}\$까지의 범위
11	Cabin	문자열 (object)	탑승객의 객실 번호 입니다.	결측값(NaN)이 매우 많습니다 (약 \$\text{77\%}\$가 결측).
12	Embarked	문자열 (object)	탑승객이 승선한 항구 입니다.	C = Cherbourg, Q = Queenstown, S = Southampton. 2개의 결측값이 존재합니다.

각 변수와 Survived와의 상관 관계 분석하기

“각 변수와 Survived와의 상관 관계를 알려주고 막대그래프로 출력해줘.”



생존(Survived)과의 상관관계 분석 결과

변수	상관계수 (Correlation)	의미
성별 (Sex, 여성=1)	0.543	가장 강한 양의 상관관계를 가집니다. 성별이 여성일수록 (\$\text{1}\$에 가까울수록) 생존율이 높습니다.
운임 (Fare)	0.257	중간 정도의 양의 상관관계. 지불한 운임이 높을수록 생존율이 높습니다.
승선 항구 (Embarked_C)	0.168	웨르부르(C)에서 승선한 탑승객의 생존율이 비교적 높습니다.
객실 등급 (Pclass)	-0.338	가장 강한 음의 상관관계를 가집니다. 등급 숫자가 높을수록(3등석일수록) 생존율이 낮습니다.
승선 항구 (Embarked_S)	-0.150	사우샘프턴(S)에서 승선한 탑승객의 생존율이 비교적 낮습니다.

 Sheets로 내보내기



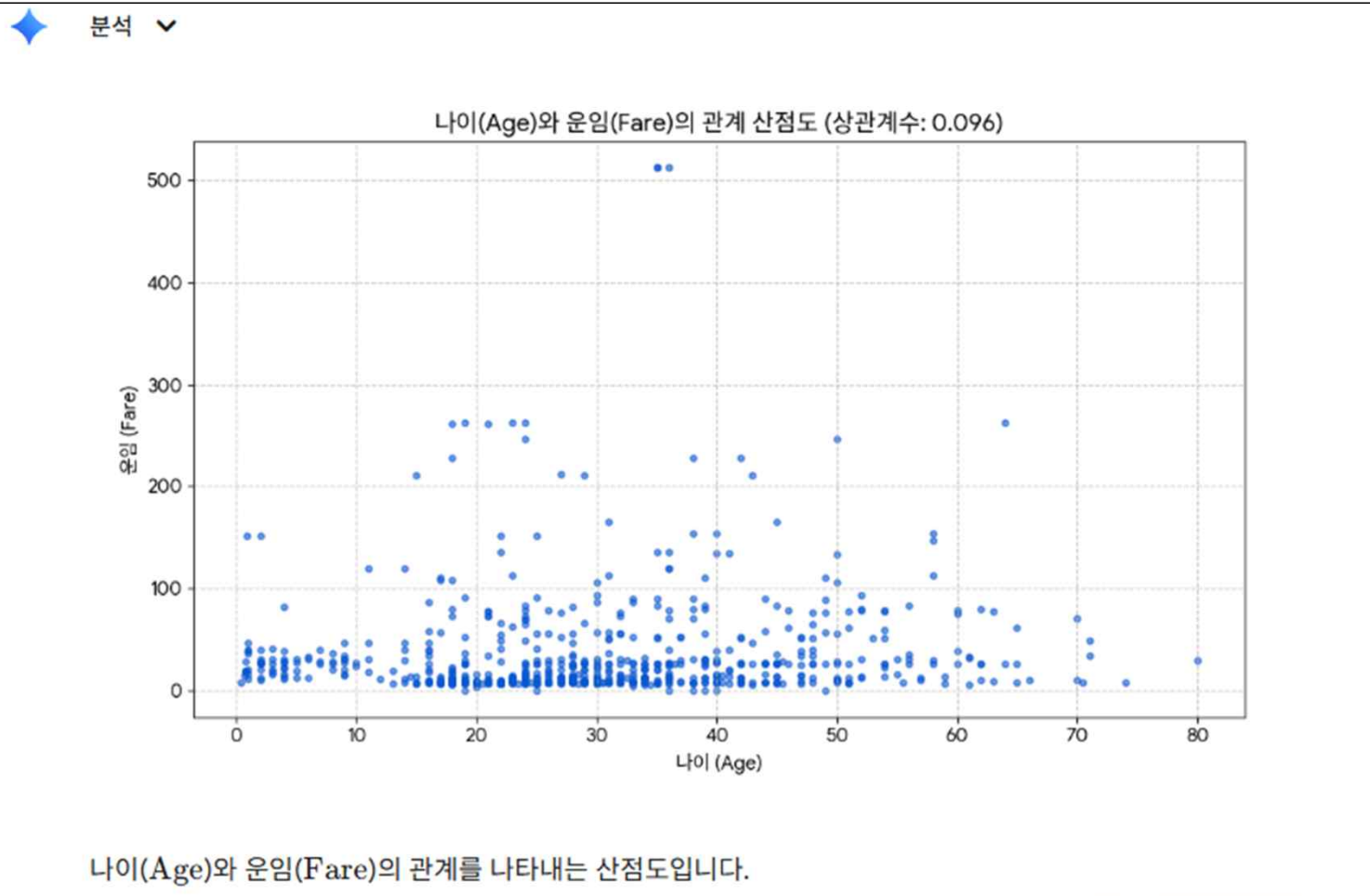
막대그래프 해석


막대그래프는 각 변수와 생존 간의 상관관계를 시각적으로 보여줍니다. 막대의 길이가 길수록 상관관계가 강하며, 파란색은 양의 상관관계(변수 값이 클수록 생존율 높음), 빨간색은 음의 상관관계(변수 값이 클수록 생존율 낮음)를 의미합니다.

1. **성별 (Sex_encoded):** 가장 강한 양의 상관관계(0.543)를 보이며, **여성이 생존에 가장 중요한 요인**이었음을 명확히 보여줍니다.
2. **객실 등급 (Pclass):** 성별 다음으로 강한 상관관계(-0.338)를 보이며, 음의 관계이므로 **객실 등급이 높을수록(1에 가까울수록) 생존율이 높았습니다.**
3. **운임 (Fare):** 양의 상관관계(0.257)는 운임이 비싼 티켓을 구매한 승객(주로 고소득층, 1등석)의 생존율이 높았음을 나타냅니다.
4. **나이 (Age):** 매우 약한 음의 상관관계(-0.077)를 보입니다.

결론적으로, ****성별(Sex)****과 ****객실 등급(Pclass)****이 타이타닉호 탑승객의 **생존에 가장 큰 영향**을 미친 변수였습니다.

“나이(Age)와 운임(fare)과의 관계를 산점도로 표현해줘”



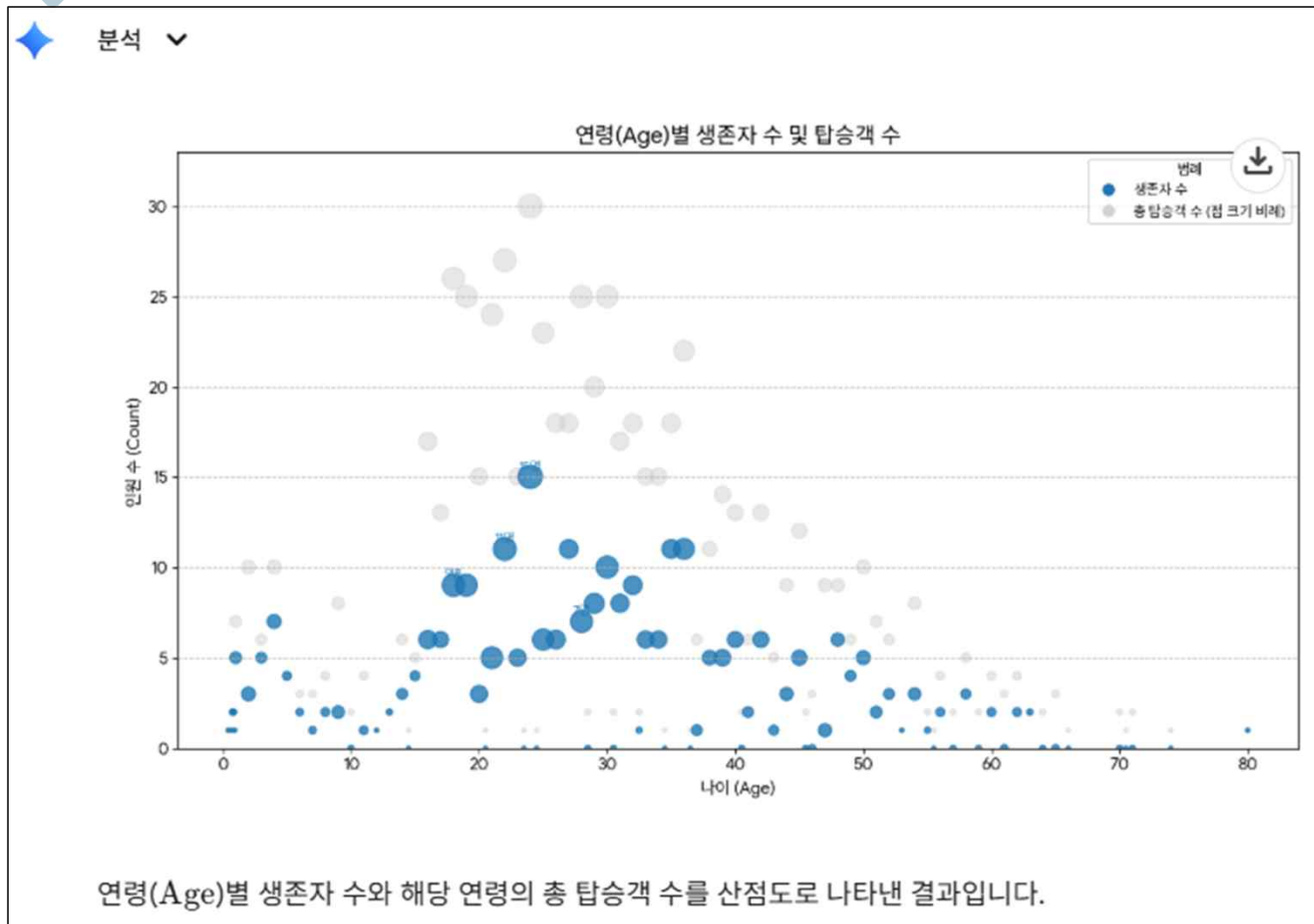


나이와 운임 관계 분석

산점도를 보면, 대부분의 탑승객은 **낮은 운임(\$100 미만)**을 지불했으며, 이 영역에서는 나이와 운임 사이에 뚜렷한 경향성을 찾기 어렵습니다.

- **상관계수:** 나이(Age)와 운임(Fare) 간의 **상관계수는 0.096**으로, 거의 관계가 없거나 매우 약한 양의 상관관계를 가집니다. 즉, 나이가 많다고 해서 운임을 더 많이 지불하거나 적게 지불하는 경향은 뚜렷하지 않습니다.
- **극단적인 운임 (Outliers):** 운임이 \$200 이상인 최고가 티켓을 구매한 탑승객은 $\text{\$}\text{20}$ 대 후반부터 $\text{\$}\text{50}$ 대 초반까지 비교적 폭넓게 분포하고 있습니다. 이들은 대부분 $\text{\$}\text{1}$ 등석 탑승객일 가능성이 높습니다.

“연령별 생존자수를 구하여 산점도로 나타내줘.”



이 그래프에서:

- **파란색 점의 높이:** 해당 연령의 생존자 수를 나타냅니다.
- **점의 크기:** 해당 연령의 총 탑승객 수에 비례하며, 점이 클수록 그 연령대의 탑승객이 많았음을 의미합니다.

연령별 생존자 수 분석

1. 가장 많은 생존자가 발생한 연령대:

- 생존자 수가 가장 많은 연령대는 $\text{\$}\text{24}\text{\$}$ 세로 $\text{\$}\text{15}\text{\$}$ 명이 생존했습니다.
- 다른 생존자 수가 많은 연령대는 $\text{\$}\text{28}\text{\$}$ 세($\text{\$}\text{14}\text{\$}$ 명), $\text{\$}\text{22}\text{\$}$ 세($\text{\$}\text{13}\text{\$}$ 명), $\text{\$}\text{36}\text{\$}$ 세($\text{\$}\text{13}\text{\$}$ 명) 등입니다.

2. 어린이 생존자:

- $\text{\$}\text{0}\text{\$}$ 세에서 $\text{\$}\text{10}\text{\$}$ 세 사이의 어린이들은 총 탑승객 수가 상대적으로 적음에도 불구하고 상당수의 생존자가 발생했습니다. 예를 들어, $\text{\$}\text{0.83}\text{\$}$ 세, $\text{\$}\text{0.92}\text{\$}$ 세, $\text{\$}\text{1}\text{\$}$ 세, $\text{\$}\text{2}\text{\$}$ 세 등 어린나이에서도 생존자($\text{\$}\text{1}\text{\$}$ 명 또는 $\text{\$}\text{2}\text{\$}$ 명)가 발생한 것을 확인할 수 있습니다.

3. 가장 많은 탑승객이 있었던 연령:

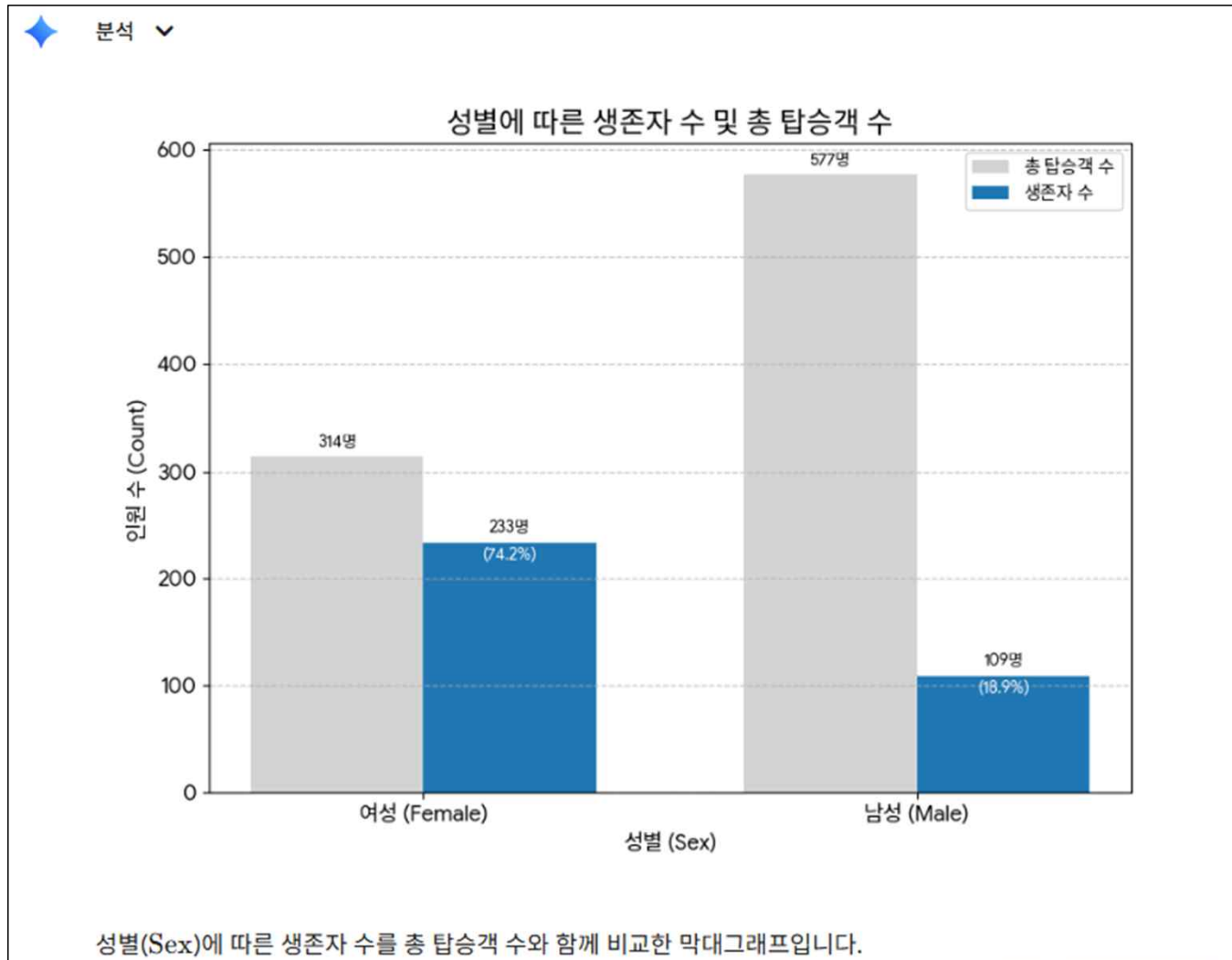
- 점의 크기를 통해 총 탑승객 수가 가장 많았던 연령대는 $\text{\$}\text{24}\text{\$}$ 세($\text{\$}\text{30}\text{\$}$ 명), $\text{\$}\text{22}\text{\$}$ 세($\text{\$}\text{27}\text{\$}$ 명), $\text{\$}\text{18}\text{\$}$ 세($\text{\$}\text{26}\text{\$}$ 명) 등입니다. 이는 젊은 성인층이 가장 많이 탑승했음을 보여줍니다.

4. 고령층 생존자:

- $\text{\$}\text{60}\text{\$}$ 세 이상의 고령층은 탑승객 수가 적지만, $\text{\$}\text{63}\text{\$}$ 세에서는 $\text{\$}\text{2}\text{\$}$ 명 중 $\text{\$}\text{2}\text{\$}$ 명 모두 생존(100% 생존)**하는 등 소수이지만 생존자가 존재했습니다.
- 가장 고령의 생존자는 $\text{\$}\text{80}\text{\$}$ 세로 $\text{\$}\text{1}\text{\$}$ 명 생존했습니다.

이 그래프는 탑승객이 많은 젊은 성인층에서 절대적인 생존자 수가 높지만, 상대적인 생존율은 나이가 매우 어린 층이나 특정 고령층에서 높았을 수 있음을 시사합니다.

“성별에 따른 생존자수를 히스토그램으로 그려줘.”



성별 생존 통계

성별	생존자 수	총 탑승객 수	생존율
여성 (Female)	$\text{\text{233}}$ 명	$\text{\text{314}}$ 명	74.2%
남성 (Male)	$\text{\text{109}}$ 명	$\text{\text{577}}$ 명	18.9%

 Sheets로 내보내기

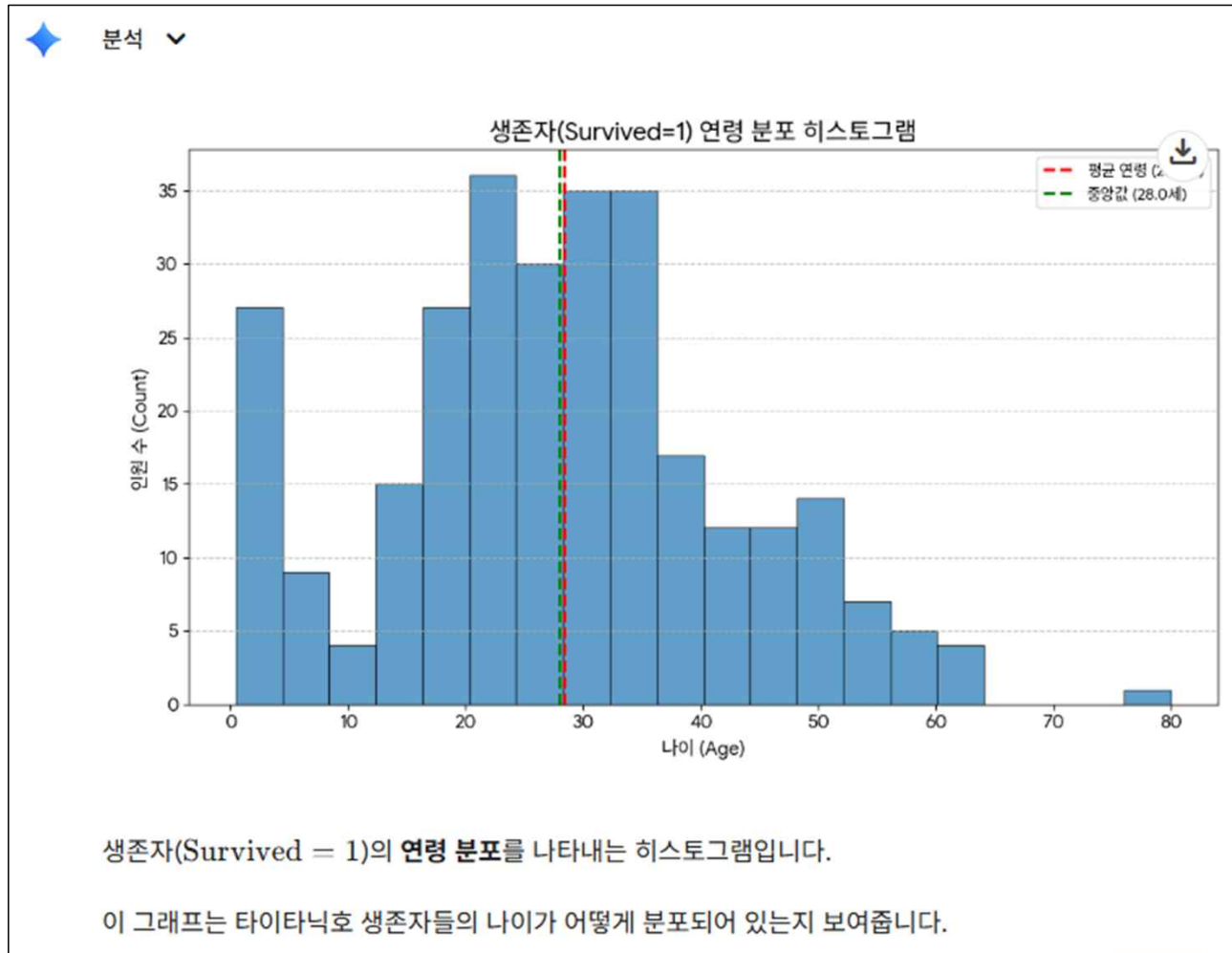


막대그래프 해석

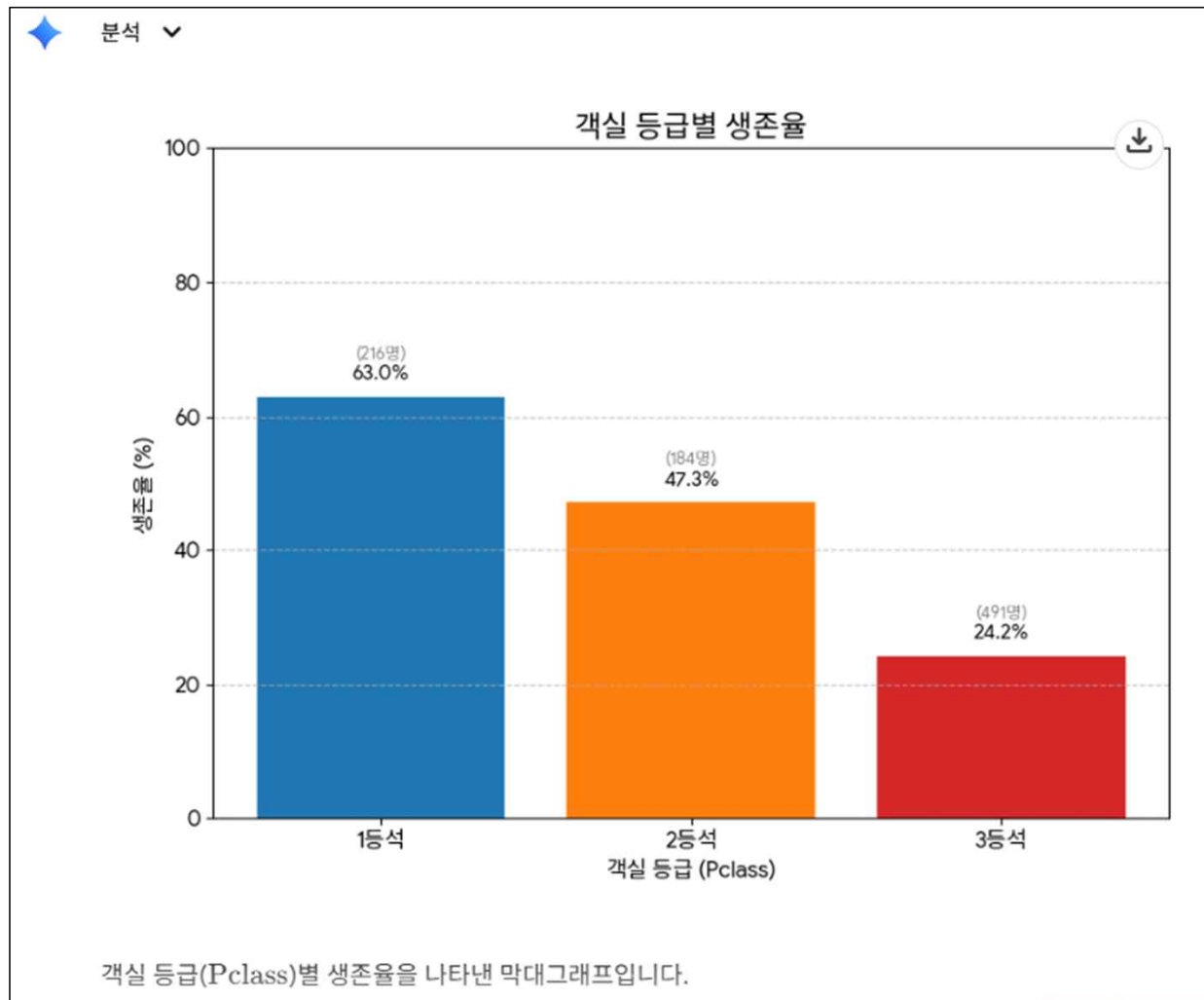
- **남성 (Male):** 총 탑승객이 $\text{\text{577}}$ 명으로 여성보다 훨씬 많았으나, 생존자는 $\text{\text{109}}$ 명에 불과하여 ****생존율이 18.9%****로 매우 낮았습니다.
- **여성 (Female):** 총 탑승객이 $\text{\text{314}}$ 명이었고, 그중 $\text{\text{233}}$ 명이 생존하여 ****생존율이 74.2%****로 압도적으로 높았습니다.

이 분석은 앞서 상관관계(0.543)에서 확인했듯이, **성별이 타이타닉호 생존에 가장 강력한 요인**이었음을 명확히 보여주며, "여성과 아이들 먼저(Women and children first)"라는 구명 원칙이 실제 생존 결과에 큰 영향을 미쳤음을 시사합니다.

“연령에 따른 생존자수를 히스토그램으로 그려줘.”



“티켓 등급별 생존율 막대그래프로 나타내줘.”



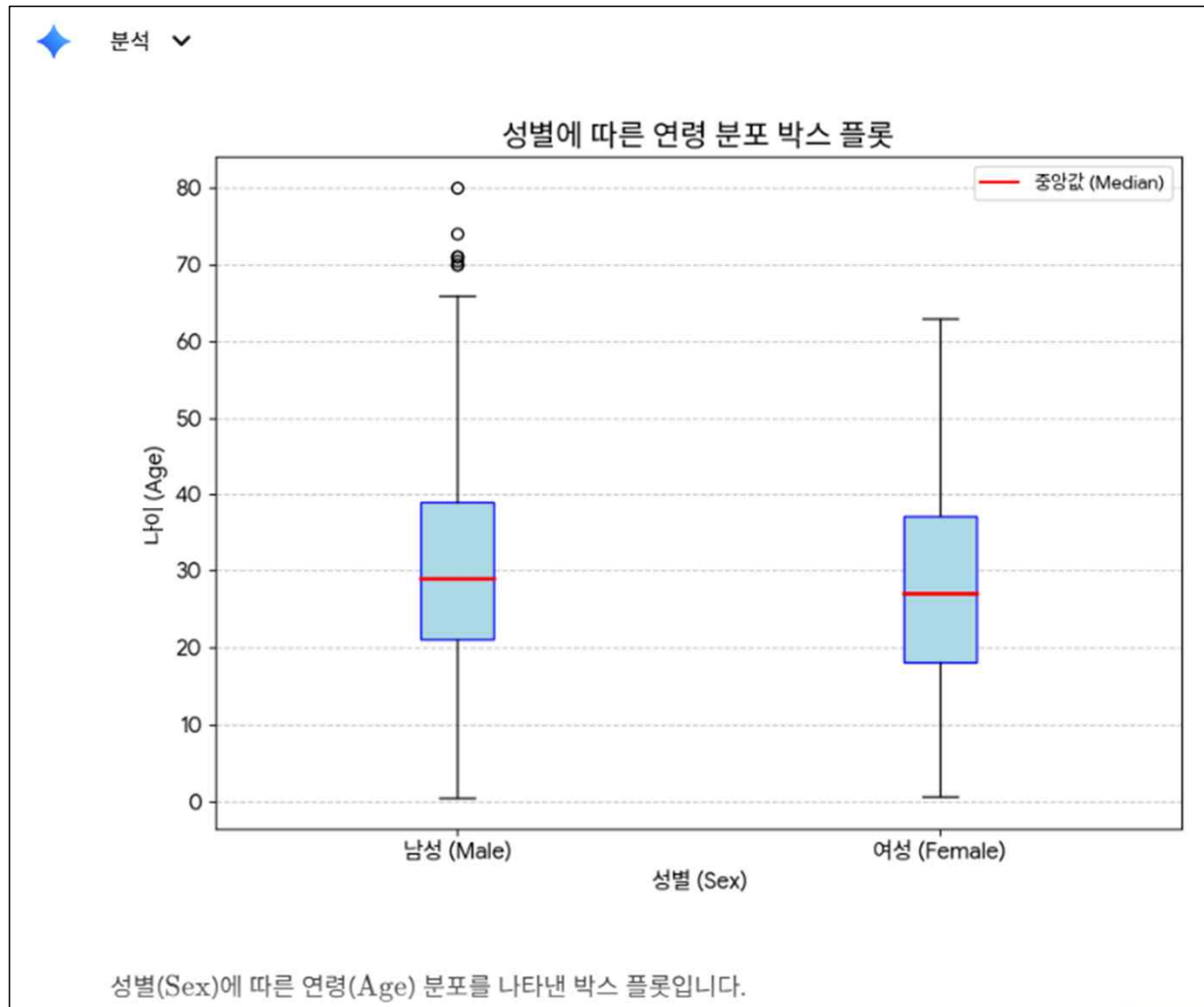
객실 등급별 생존율 분석

객실 등급	총 탑승객 수	생존율
1등석	216명	63.0%
2등석	184명	47.3%
3등석	491명	24.2%

막대그래프 해석

1. **객실 등급이 높을수록 생존율이 높습니다:** 객실 등급은 생존율과 강한 상관관계(음의 상관계수 -0.338)를 가짐을 재확인합니다.
 - **1등석 (1st Class):** 63.0\% 의 가장 높은 생존율을 기록했습니다.
 - **3등석 (3rd Class):** 491 명으로 가장 많은 탑승객이 있었음에도 불구하고, 생존율은 24.2\% 로 가장 낮았습니다.
2. **생존 격차:** 1 등석 생존율(63.0\%)은 3 등석 생존율(24.2\%)보다 2.5 배 이상 높았습니다. 이는 타이타닉호의 비상 상황에서 **경제적 지위**가 생존 가능성에 중대한 영향을 미쳤음을 보여줍니다.

“연령 분포를 성별로 함께 박스 플롯으로 그려줘.”



박스 플롯 해석:

1. 중앙값(Median) 비교:

- **남성의 중앙 연령(29.0 세)이 여성의 중앙 연령(27.0 세)보다 약간 더 높았습니다. 즉, 남성 탑승객이 여성 탑승객보다 평균적으로 아주 조금 더 나이가 많았습니다.**

2. 연령대 분포(IQR) 비교:

- IQR (박스)는 데이터의 중간 50\% 가 분포하는 범위를 나타냅니다.
- **여성** 탑승객의 중간 50\% 는 18 세에서 37 세 사이에 분포했습니다.
- **남성** 탑승객의 중간 50\% 는 21 세에서 39 세 사이에 분포했습니다.
- 두 성별 모두 연령 분포의 범위(IQR)는 18 세와 19 세로 유사했습니다.

3. 최솟값과 최댓값 (수염):

- 두 성별 모두 **수염(Whiskers)**이 아래쪽으로 길게 뻗어 있습니다. 이는 어린 나이의 탑승객(어린이)들이 있었음을 보여줍니다.
- **최댓값(Max):** 남성 탑승객 중에는 80 세의 최고령자가 있었습니다.

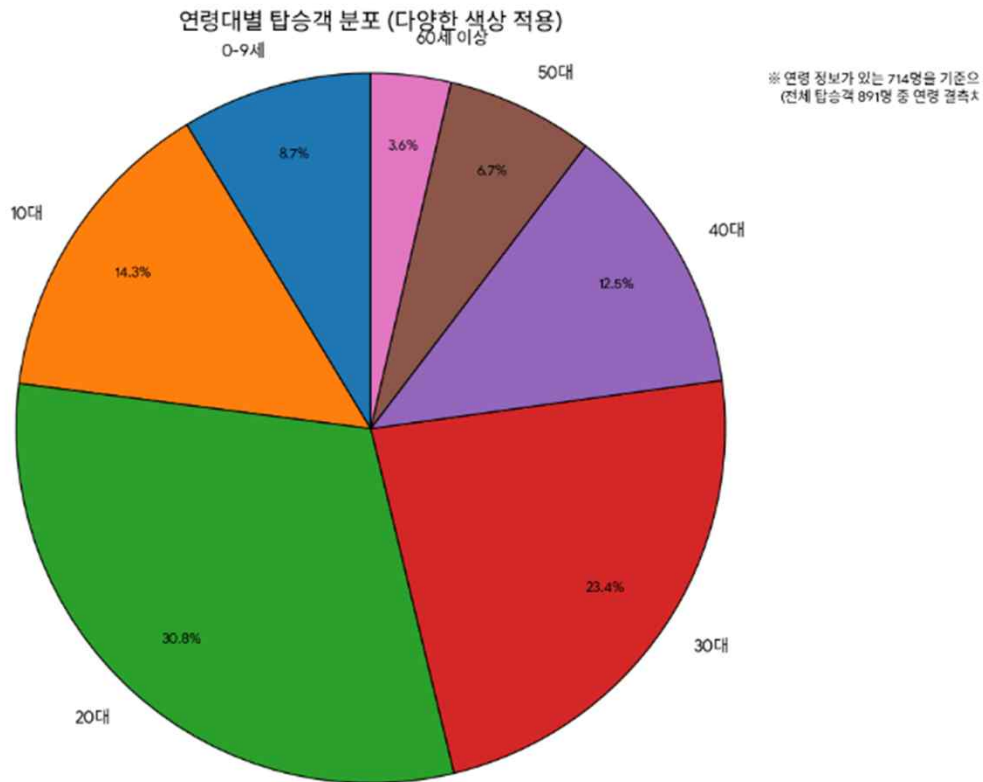
4. 이상치 (Outliers, 개별 점):

- 남성과 여성 모두 60 대 중반 이상의 고령층에서 몇몇 이상치가 확인됩니다. 이는 해당 연령대 탑승객이 소수였음을 의미합니다.

결론적으로, 타이타닉호 탑승객의 연령 분포는 성별에 따라 큰 차이를 보이지 않았으며, 남성 중앙값이 여성보다 2 세 정도 높았습니다.

“연령별 분포를 파이차트로 그려줘(색상을 다양하게 사용해줘).”

◆ 분석 ▾



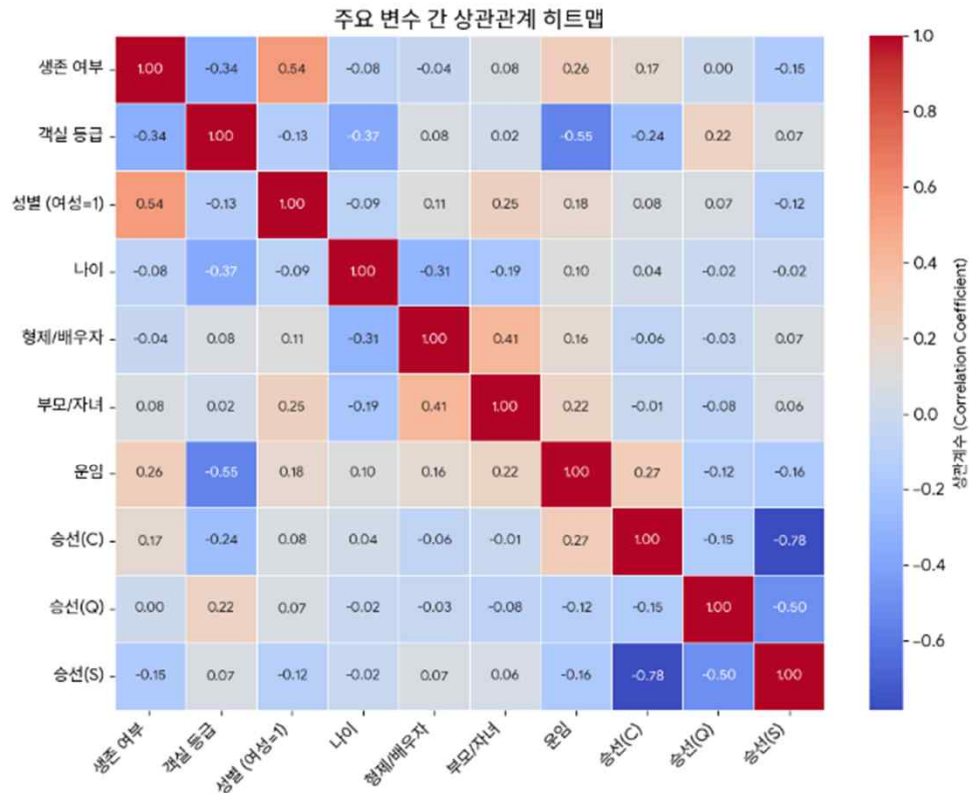
연령대별 탑승객 분포를 다양한 색상으로 구분하여 나타낸 파이 차트입니다.

연령대별 탑승객 분포 요약 (총 \$\text{714}\$명 기준)

연령대	인원 수	비율
20대 (\$\text{20-29}\$세)	\$\text{220}\$명	30.8%
30대 (\$\text{30-39}\$세)	\$\text{167}\$명	23.4%
40대 (\$\text{40-49}\$세)	\$\text{89}\$명	12.5%
50대 (\$\text{50-59}\$세)	\$\text{48}\$명	6.7%
60세 이상	\$\text{26}\$명	3.6%
0-9세 (어린이)	\$\text{62}\$명	8.7%
10대 (\$\text{10-19}\$세)	\$\text{102}\$명	14.3%

"각 변수와 Survived와의 상관 관계를 히트맵으로 그려줘."

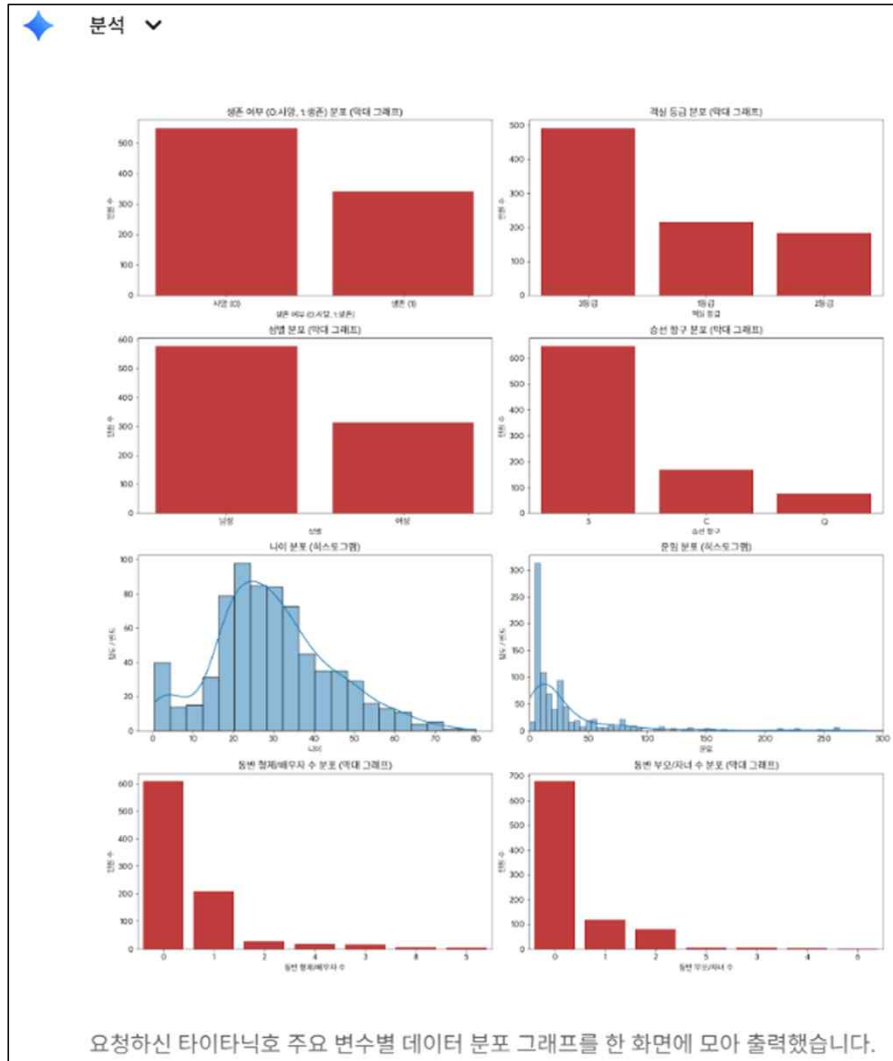
◆ 분석 ▼



타이타닉호의 주요 변수 간 상관관계를 나타낸 히트맵입니다.

히트맵은 변수 쌍 사이의 선형 관계를 보여주며, **빨간색**은 강한 양의 상관관계(두 변수가 함께 증가/감소), **파란색**은 강한 음의 상관관계(한 변수가 증가할 때 다른 변수가 감소)를 의미합니다.

“각 변수별 데이터 분포 그래프를 한 화면에 출력해줘.”



주요 변수별 데이터 분포 해석

1. 범주형/이산형 변수 (막대 그래프)

- 생존 여부 (Survived): 사망(0)한 인원이 생존(1)한 인원보다 훨씬 많습니다.
- 객실 등급 (Pclass): $\text{\$}\text{text}\{3\}$ 등석 탑승객이 다른 등급($\text{\$}\text{text}\{1\}$ 등석, $\text{\$}\text{text}\{2\}$ 등석)보다 압도적으로 많습니다.
- 성별 (Sex): 남성 탑승객이 여성 탑승객보다 약 $\text{\$}\text{text}\{2\}$ 배가량 많습니다.
- 승선 항구 (Embarked): **S (Southampton)**에서 탑승한 인원이 가장 많습니다.
- 동반 형제/배우자 (SibSp) 및 부모/자녀 (Parch): 두 변수 모두 **0 (동반 가족 없음)**인 경우가 대다수를 차지합니다.

2. 연속형 변수 (히스토그램)

- 나이 (Age):
 - $\text{\$}\text{text}\{20\}$ 대 초반에서 $\text{\$}\text{text}\{30\}$ 대 초반에 탑승객이 가장 집중되어 있습니다.
 - $\text{\$}\text{text}\{0\}$ 세 근처에서 또 하나의 작은 봉우리가 관찰되며, 이는 어린이 탑승객을 나타냅니다.
 - 전반적으로 오른쪽으로 꼬리가 긴 형태(약간의 오른쪽 편포)를 보입니다.
- 운임 (Fare):
 - 대부분의 탑승객이 $\text{\$}50$ 미만의 낮은 운임을 지불했습니다 (그래프에서 가장 왼쪽 봉우리).
 - 운임이 높은 소수의 이상치(Outlier) 때문에 분포가 매우 심하게 오른쪽으로 편향되어 있습니다. (그래프는 가독성을 위해 $\text{\$}\text{text}\{\$300\}$ 까지만 표시되었습니다.)

이 그래프들을 통해 데이터의 전반적인 형태와 분포의 불균형을 한눈에 파악할 수 있습니다.



감사합니다