

데이터 시각화

1. 시각적 탐색

시각적 탐색

시각적 탐색은 데이터를 시각적으로 표현하여 데이터의 다양한 특성을 이해하고 분석하는 과정이다. 이는 단순히 숫자나 텍스트 형태로 데이터를 보는 것보다 훨씬 직관적이고 효과적인 분석 방법이다. 시각적 탐색을 통해 데이터의 구조, 패턴, 트렌드, 이상치 등을 쉽게 발견할 수 있다.

1. 시각적 탐색

시각적 탐색의 중요성

- **빠른 이해:** 시각적 표현은 복잡한 데이터를 빠르게 이해할 수 있게 도와준다. 사람의 뇌는 시각 정보를 처리하는 데 훨씬 능숙하다.
- **패턴 인식:** 데이터의 시각화를 통해 숨겨진 패턴이나 트렌드를 쉽게 식별할 수 있다. 이는 중요한 인사이트를 얻는 데 필수적이다.
- **이상치 탐지:** 시각화 된 데이터는 이상치를 쉽게 식별할 수 있게 도와준다. 이상치는 데이터 분석에 중요한 영향을 미칠 수 있으므로 이를 식별하는 것이 중요하다.
- **커뮤니케이션:** 데이터를 시각적으로 표현하면 다른 사람과 데이터를 공유하고 논의하는 과정에서 효과적이다. 그래프나 차트는 복잡한 데이터를 간결하고 명확하게 전달할 수 있다.

1. 시각적 탐색

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) 막대 그래프 (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.

1. 시각적 탐색

시각적 탐색의 과정

- 1. 데이터 준비:** 시각화를 위해 데이터를 정리하고 정제한다. 데이터의 정확성과 일관성을 확보하는 단계이다.
- 2. 그래프 선택:** 분석하고자 하는 데이터의 특성에 맞는 그래프나 차트를 선택한다. 예를 들어, 두 변수 간의 관계를 분석할 때는 산점도를, 범주형 데이터의 빈도를 비교할 때는 막대그래프를 선택한다.
- 3. 시각화 도구 사용:** Python의 Matplotlib, Seaborn, R의 ggplot2, Excel 등 다양한 시각화 도구를 사용하여 데이터를 시각화한다.
- 4. 해석 및 인사이트 도출:** 시각화된 데이터를 해석하고, 이를 통해 의미 있는 인사이트를 도출한다. 이 과정에서 데이터의 패턴, 트렌드, 이상치를 분석한다.

시각적 탐색의 사례

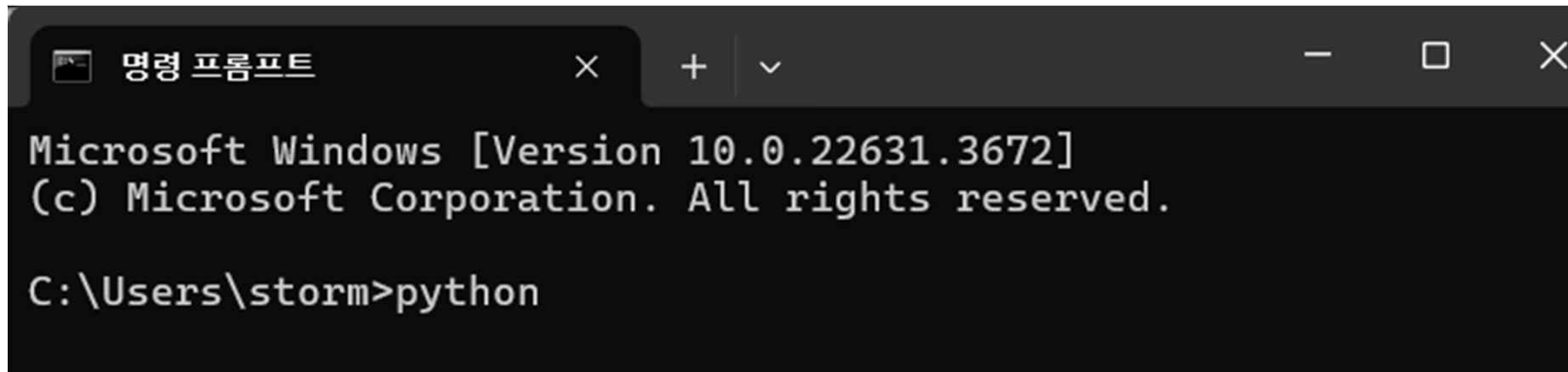
- 1. 매출 분석:** 매출 데이터를 월별로 시각화하여 매출의 증가 또는 감소 추세를 파악하고, 특정 시점의 이상치를 식별할 수 있다.
- 2. 고객 분석:** 고객의 연령대, 성별, 지역별 분포를 시각화하여 마케팅 전략을 수립할 수 있다.
- 3. 주식 시장 분석:** 주식 가격의 시간에 따른 변화를 시각화하여 시장의 트렌드와 변동성을 분석할 수 있다.

시각적 탐색은 데이터 분석의 필수적인 단계로, 데이터를 보다 깊이 이해하고 효과적인 의사 결정을 내리는 데 중요한 역할을 한다. 이를 통해 데이터에서 더 많은 가치를 창출할 수 있다.

1. 시각적 탐색

Python 설치하기

Windows의 명령 프롬프트에서 python 명령 실행



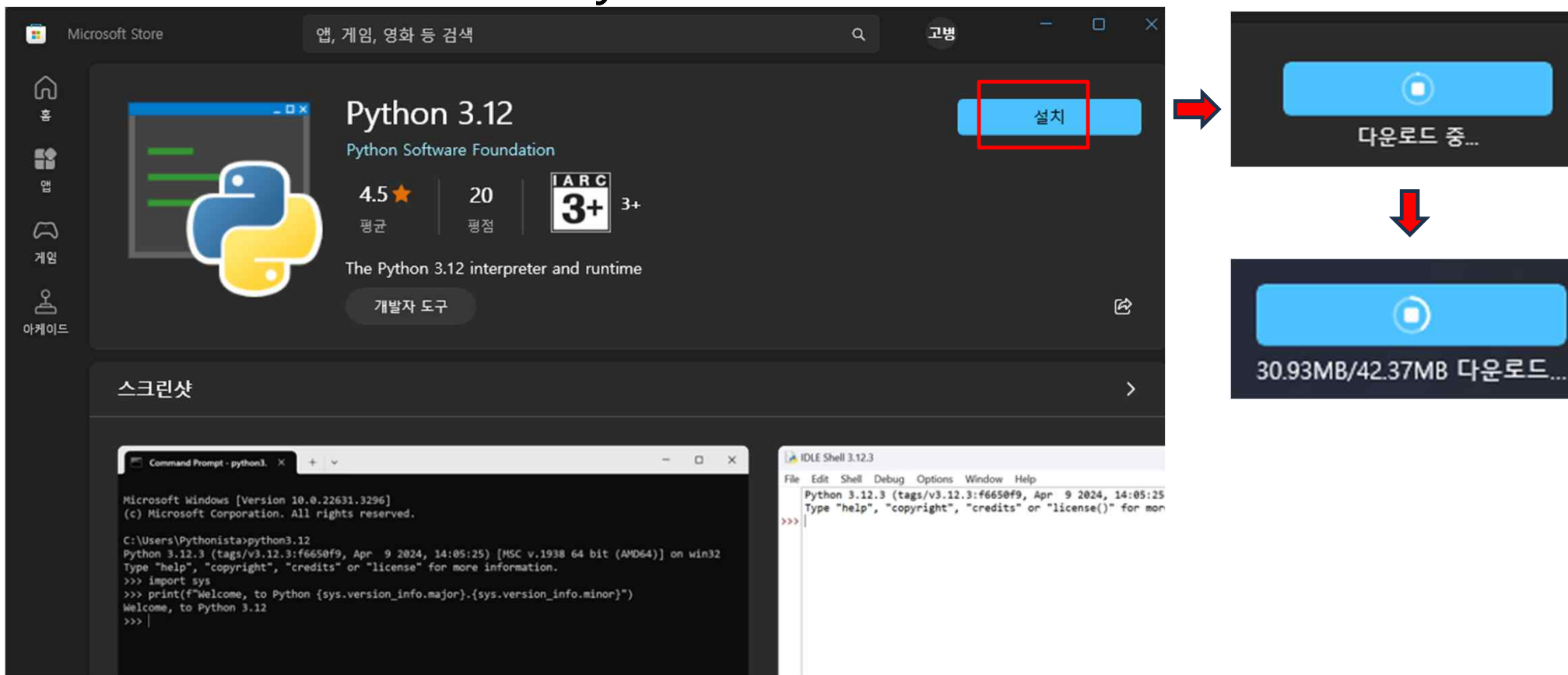
```
명령 프롬프트
Microsoft Windows [Version 10.0.22631.3672]
(c) Microsoft Corporation. All rights reserved.

C:\Users\storm>python
```

1. 시각적 탐색

Python 설치하기

Microsoft Store에서 Python 3.12 설치



1. 시각적 탐색

설치 완료후 명령 프롬프트에서 그래픽 라이브러리 설치
pip install pandas matplotlib seaborn scipy

```
명령 프롬프트 - cmd
Microsoft Windows [Version 10.0.22631.3672]
(c) Microsoft Corporation. All rights reserved.

C:\Users\storm>cmd
Microsoft Windows [Version 10.0.22631.3672]
(c) Microsoft Corporation. All rights reserved.

C:\Users\storm>pip install pandas matplotlib seaborn scipy
Defaulting to user installation because normal site-packages is not writeable
Collecting pandas
  Downloading pandas-2.2.2-cp312-cp312-win_amd64.whl.metadata (19 kB)
Collecting matplotlib
  Downloading matplotlib-3.9.0-cp312-cp312-win_amd64.whl.metadata (11 kB)
Collecting seaborn
  Using cached seaborn-0.13.2-py3-none-any.whl.metadata (5.4 kB)
Collecting scipy
  Downloading scipy-1.13.1-cp312-cp312-win_amd64.whl.metadata (60 kB)
  60.6/60.6 kB 3.4 MB/s eta 0:00:00
Collecting numpy>=1.26.0 (from pandas)
  Downloading numpy-1.26.4-cp312-cp312-win_amd64.whl.metadata (61 kB)
  61.0/61.0 kB 3.4 MB/s eta 0:00:00
```

1. 시각적 탐색

Python 설치 확인

Windows의 명령 프롬프트에서 python 명령 실행


```
python  
import matplotlib
```

```
C:\Users\storm>python  
Python 3.12.3 (tags/v3.12.3:f6650f9, Apr 9 2024, 14:05:25) [MSC v.1938 64 bit (AMD64)] on win32  
Type "help", "copyright", "credits" or "license" for more information.  
>>> import matplotlib  
>>>
```

1. 시각적 탐색

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](#)

1. 시각적 탐색

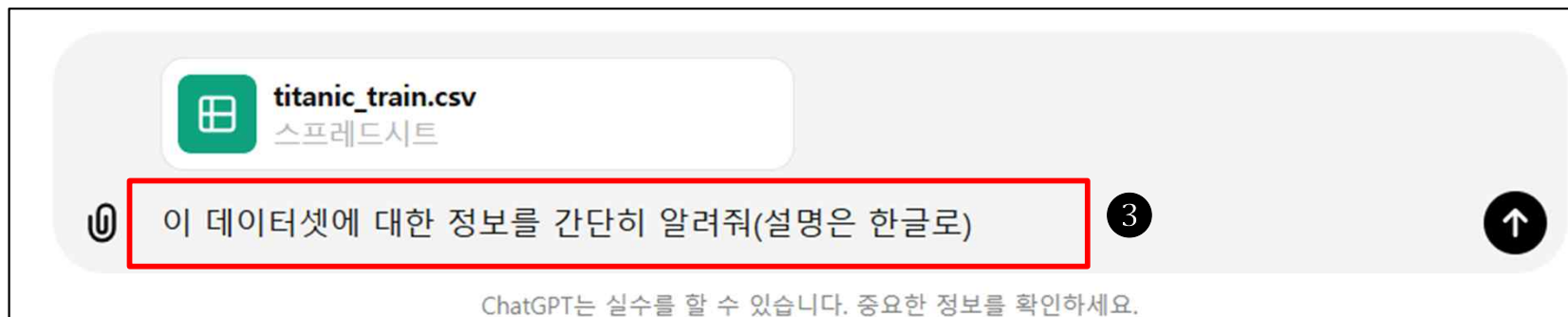
titanic_train.csv 파일 엑셀에서 확인(원본 파일명은 train.csv)

	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, M	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, female		38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen female		26	0	0	STON/O2.	7.925		S
5	4	1	1	Futrelle, M	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, M	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, male		54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, M	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstrom female		4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S
14	13	0	3	Saunders, male		20	0	0	A/5. 2151	8.05		S
15	14	0	3	Andersson, male		39	1	5	347082	31.275		S
16	15	0	3	Vestrom, M	female	14	0	0	350406	7.8542		S
17	16	1	2	Hewlett, M	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, M	male		0	0	244373	13		S
20	19	0	3	Vander Planck female		31	1	0	345763	18		S
21	20	1	3	Masella female			0	0	2649	7.225		C
22	21	0	2	Fynney, M	male	35	0	0	239865	26		S
23	22	1	2	Beesley, M	male	34	0	0	248698	13	D56	S
24	23	1	3	McGowan female		15	0	0	330923	8.0292		Q
25	24	1	1	Sloper, M	male	28	0	0	113788	35.5	A6	S
26	25	0	3	Palsson, M	female	8	3	1	349909	21.075		S

1. 시각적 탐색

Titanic 생존자 데이터셋 사용

- 데이터셋 로드 및 확인: 데이터셋을 불러오고 기본 정보를 확인.



1. 시각적 탐색

Titanic Train					↓ ↗
	PassengerId	Survived	Pclass	Name	
1	1	0	3	Braund, Mr. Owen Harris	
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	
3	3	1	3	Heikkinen, Miss. Laina	
4	4	1	1	Futrelle, Mrs. Jacques Heath	

이 데이터셋에 대한 정보를 간단히 알려줘(설명은 한글로)

1. 시각적 탐색



이 데이터셋은 타이타닉 호 승객에 대한 정보입니다. 총 891명의 승객 데이터가 포함되어 있습니다. 각 승객에 대해 12개의 열(column)이 있으며, 각 열은 다음과 같은 정보를 담고 있습니다:

1. **PassengerId**: 승객 ID
2. **Survived**: 생존 여부 (0: 사망, 1: 생존)
3. **Pclass**: 티켓 등급 (1: 1등급, 2: 2등급, 3: 3등급)
4. **Name**: 승객 이름
5. **Sex**: 성별
6. **Age**: 나이
7. **SibSp**: 함께 탑승한 형제자매 또는 배우자 수
8. **Parch**: 함께 탑승한 부모 또는 자녀 수
9. **Ticket**: 티켓 번호
10. **Fare**: 요금
11. **Cabin**: 객실 번호
12. **Embarked**: 탑승한 항구 (C: Cherbourg, Q: Queenstown, S: Southampton)

1. 시각적 탐색

특징적인 점으로는 **Age** 열에 결측값이 있으며, **Cabin** 열에 많은 결측값이 존재합니다. **Embarked** 열에도 소수의 결측값이 있습니다.

데이터의 첫 5행은 다음과 같습니다:

- **PassengerId 1:** Owen Harris Braund, 남성, 22세, 3등급 티켓, 요금 7.25
- **PassengerId 2:** Florence Briggs Thayer Cumings, 여성, 38세, 1등급 티켓, 요금 71.2833
- **PassengerId 3:** Laina Heikkinen, 여성, 26세, 3등급 티켓, 요금 7.925
- **PassengerId 4:** Lily May Peel Futrelle, 여성, 35세, 1등급 티켓, 요금 53.1
- **PassengerId 5:** William Henry Allen, 남성, 35세, 3등급 티켓, 요금 8.05

이 데이터는 생존 예측 모델을 만들거나, 타이타닉 승객의 특징을 분석하는 데 유용하게 사용할 수 있습니다. [↗]

1. 시각적 탐색

각 변수와 Survived와의 상관 관계 분석하기

“각 변수와 Survived와의 상관 관계를 파악해줘.”



타이타닉 데이터셋에서 각 변수와 생존 여부(Survived) 간의 상관 관계는 다음과 같습니다:

- **Pclass** (티켓 등급): -0.338481 (부의 상관 관계, 티켓 등급이 높을수록 생존 확률이 높음)
- **Sex** (성별): 0.543351 (양의 상관 관계, 여성이 생존 확률이 높음)
- **Age** (나이): -0.077221 (약한 부의 상관 관계)
- **SibSp** (형제자매/배우자 수): -0.035322 (거의 상관 관계 없음)
- **Parch** (부모/자녀 수): 0.081629 (약한 양의 상관 관계)
- **Fare** (요금): 0.257307 (양의 상관 관계, 요금이 높을수록 생존 확률이 높음)
- **PassengerId** (승객 ID): -0.005007 (거의 상관 관계 없음)

이 상관 관계는 변수들이 생존 여부에 미치는 영향을 이해하는 데 중요한 정보를 제공합니다. 성별 (Sex)과 티켓 등급(Pclass), 요금(Fare)이 생존 여부와 비교적 강한 상관 관계를 보입니다. [↗]

1. 시각적 탐색

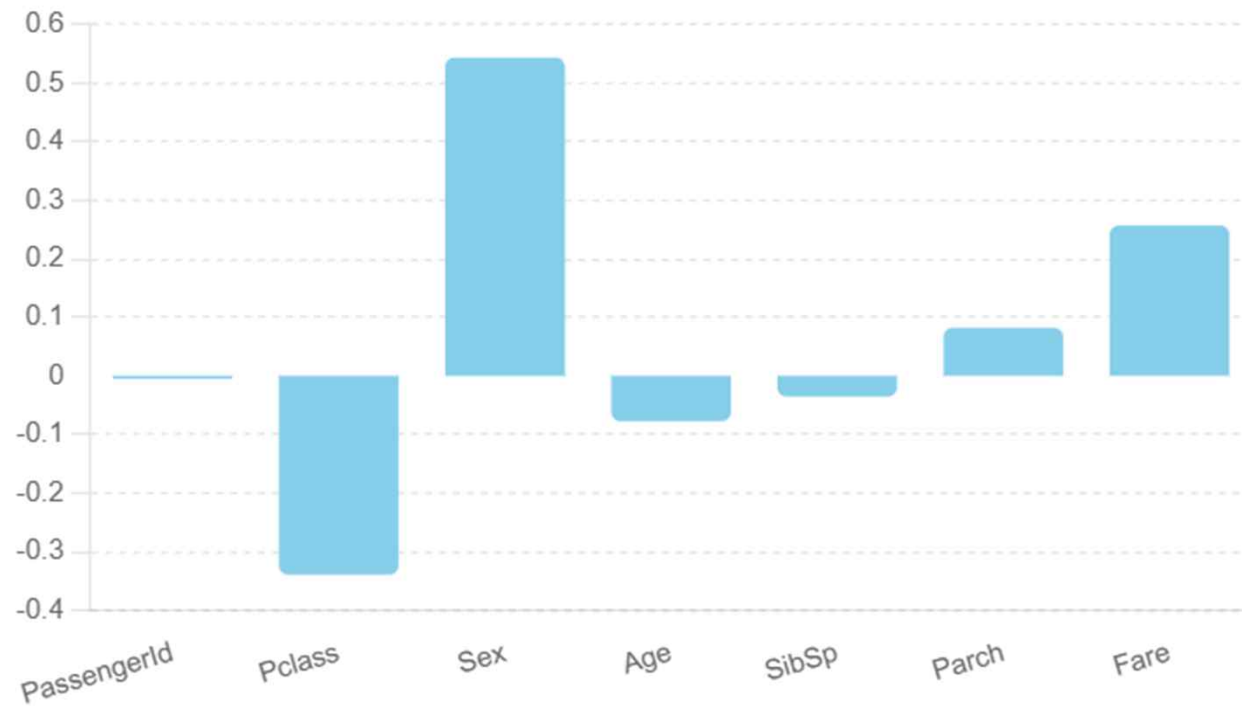
“막대 그래프로 나타내줘”



Correlation Of Variables With Survived



Y Correlation Coefficient X Variables



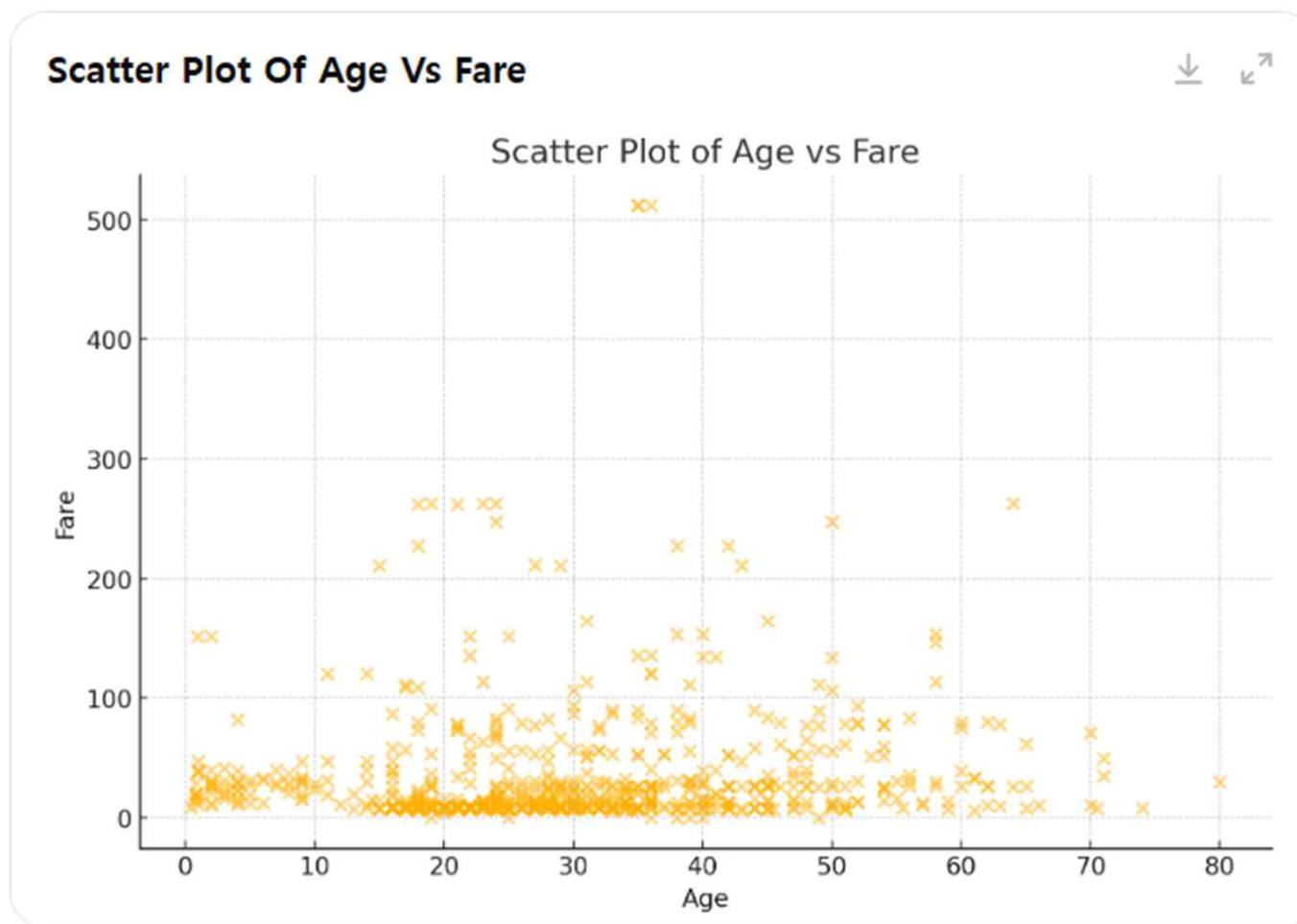
2. 데이터 시각화 그래프 종류

1. 산점도 (Scatter Plot)

- **특징:** 두 변수 간의 관계를 나타내는 데 사용된다.
- **용도:** 상관 관계를 시각적으로 파악할 때 유용하다. 예를 들어, 나이와 수입 간의 관계, 체중과 키의 관계 등을 분석할 때 사용된다.
- **장점:** 데이터의 분포와 상관성을 직관적으로 이해할 수 있다.
- **단점:** 많은 데이터 포인트가 겹치면 해석이 어려울 수 있다.

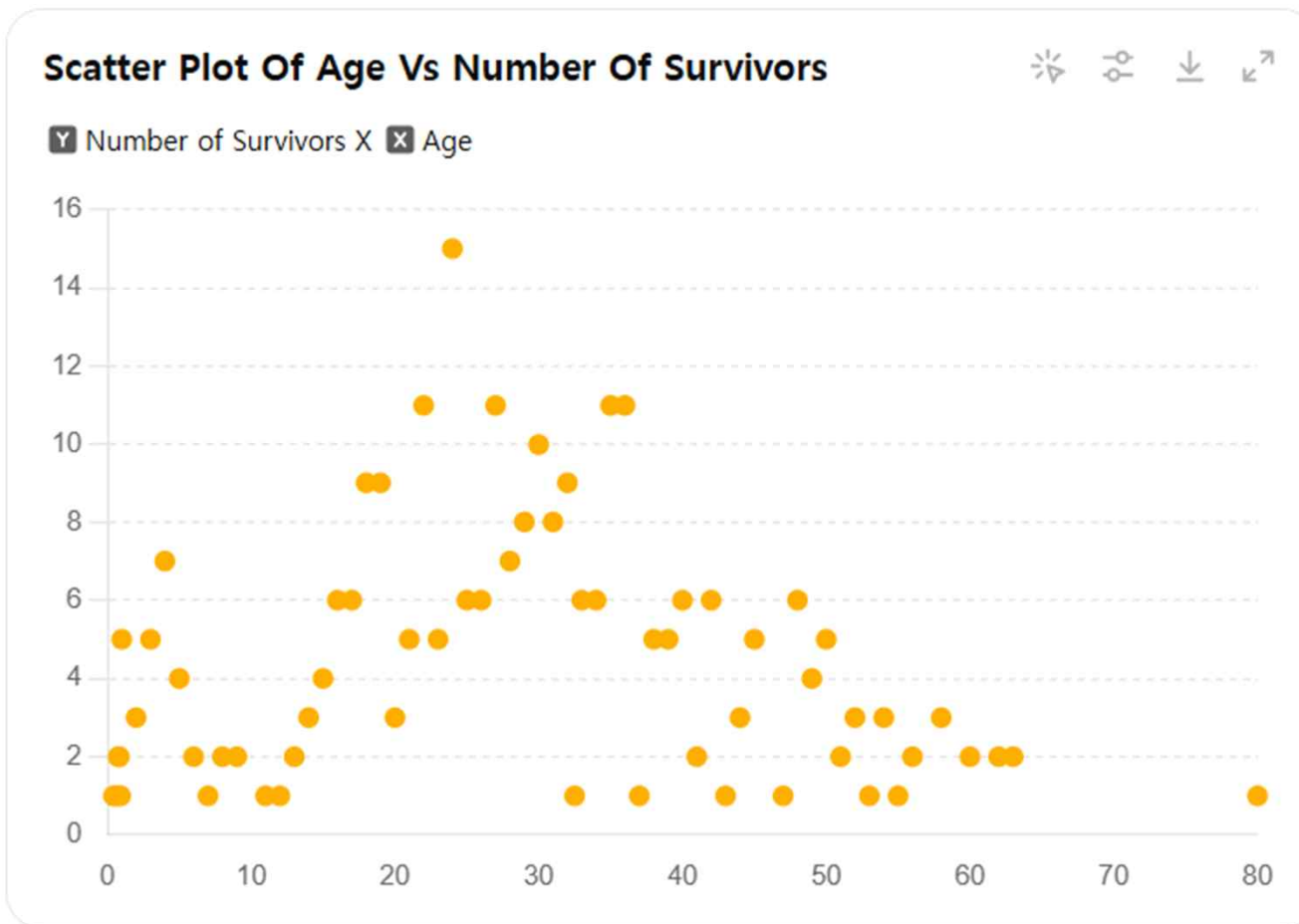
2. 데이터 시각화 그래프 종류

나이(Age)와 운임(fare)과의 관계를 산점도로 표현해줘



2. 데이터 시각화 그래프 종류

“연령에 따른 생존자수를 산점도로 나타내줘.”

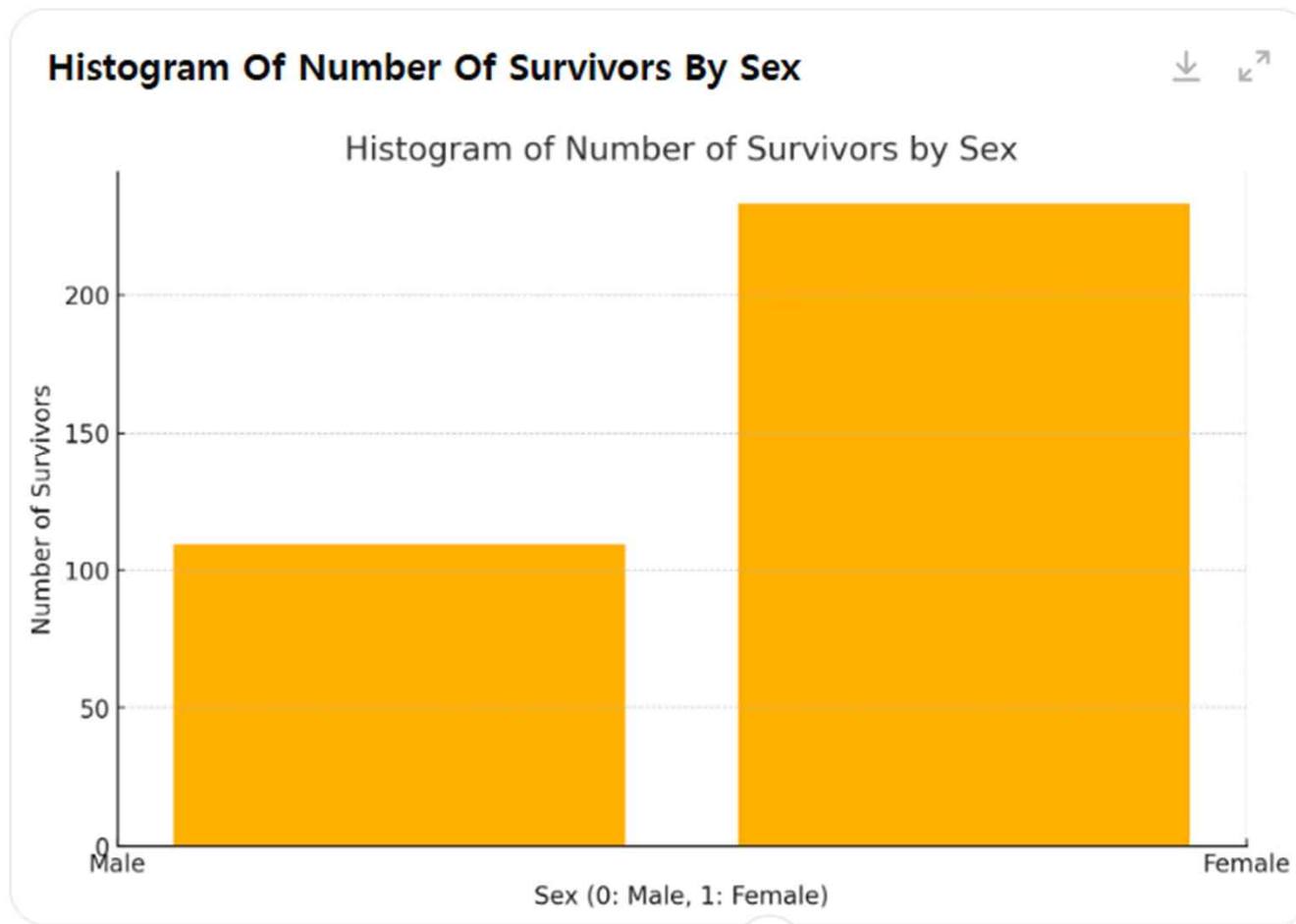


2. 히스토그램 (Histogram)

- **특징:** 연속형 데이터의 분포를 나타낸다.
- **용도:** 데이터의 분포, 중심 경향, 변동성을 파악할 때 유용하다. 예를 들어, 시험 점수의 분포, 나이 분포 등을 분석할 때 사용된다.
- **장점:** 데이터의 분포와 빈도를 한눈에 볼 수 있다.
- **단점:** 빈(bin, 구간) 크기에 따라 결과가 달라질 수 있어 주의가 필요하다.

2. 데이터 시각화 그래프 종류

성별에 따른 생존자수를 히스토그램으로 그려줘.

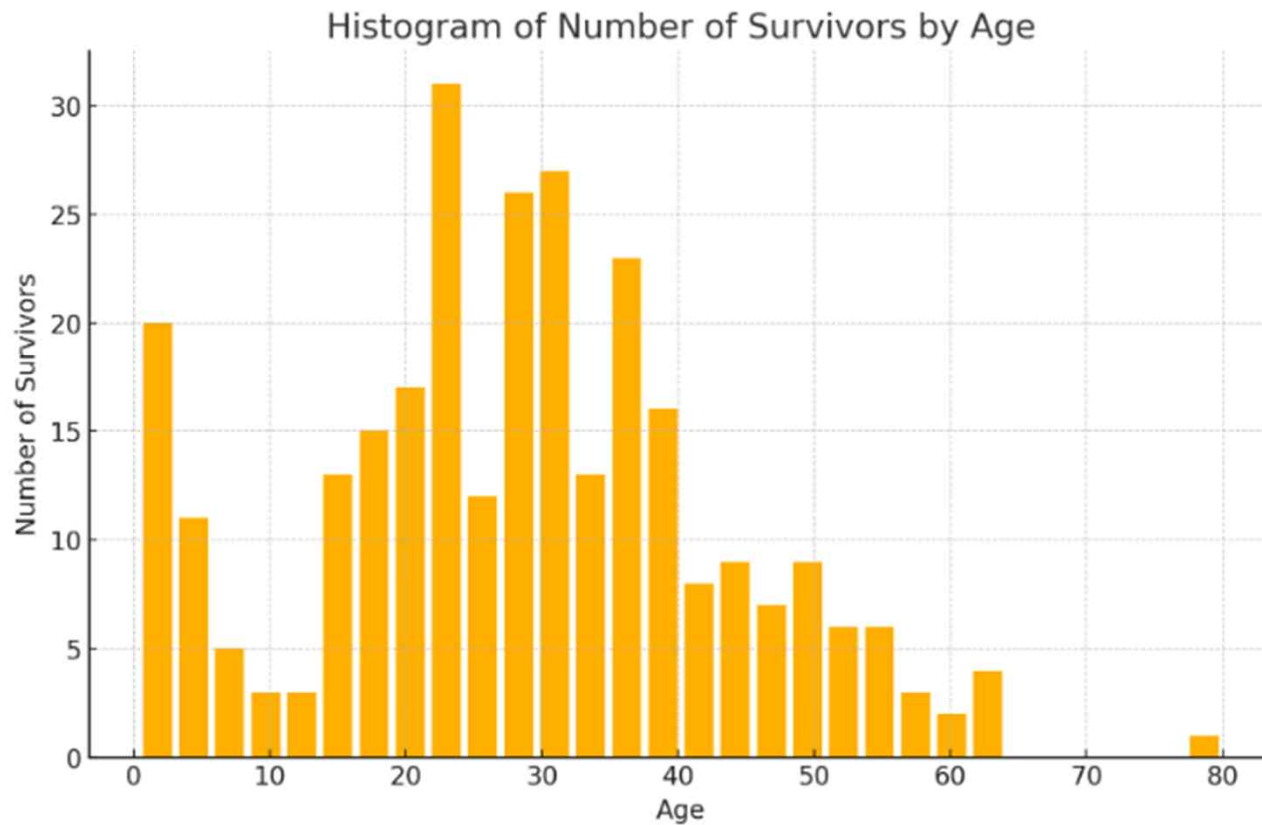


2. 데이터 시각화 그래프 종류

연령에 따른 생존자수를 히스토그램으로 그려줘.



Histogram Of Number Of Survivors By Age

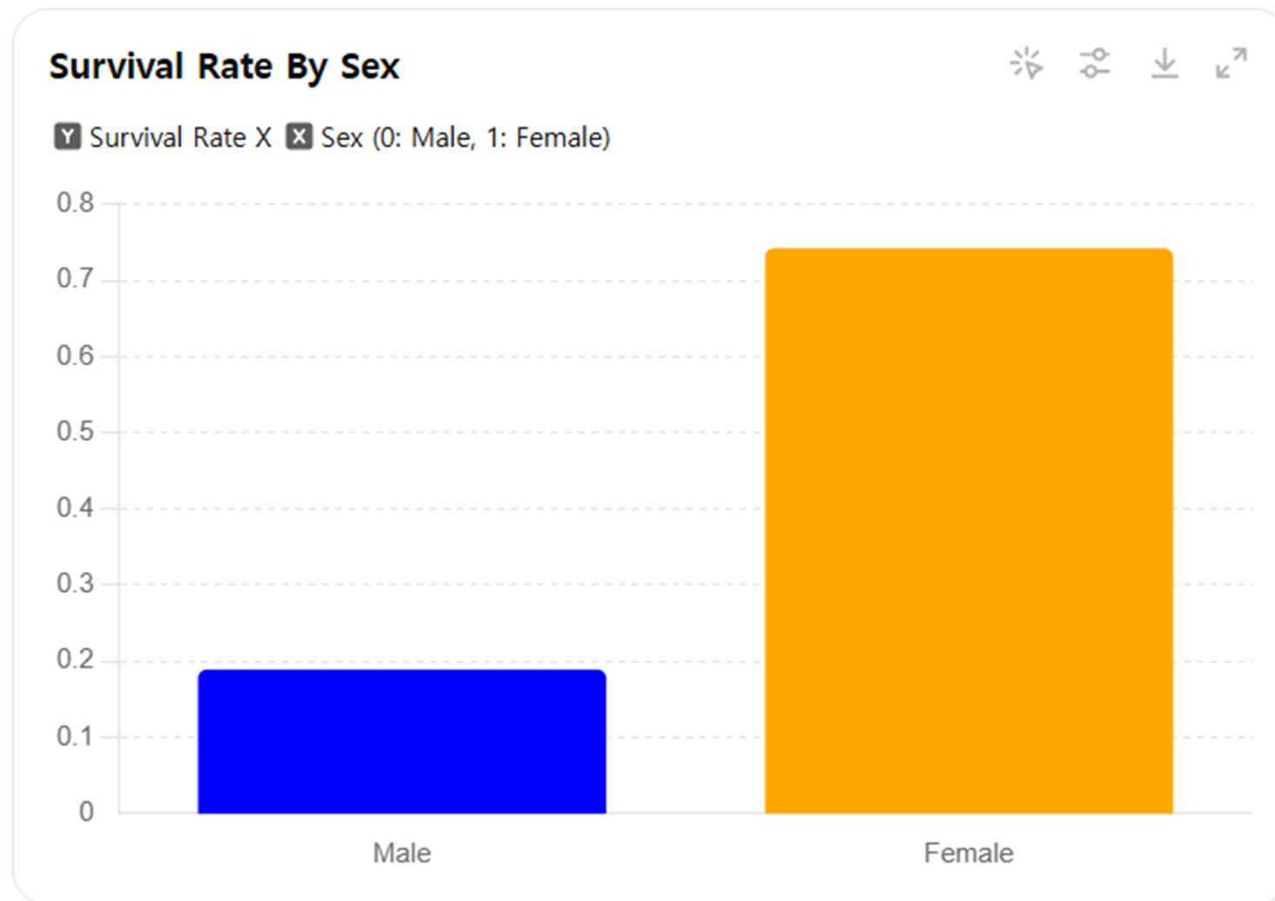


3. 막대그래프 (Bar Chart)

- **특징:** 범주형 데이터의 빈도나 비율을 나타낸다.
- **용도:** 여러 카테고리의 값을 비교할 때 유용하다. 예를 들어, 각 연도별 판매량, 각 지역별 인구 수 등을 비교할 때 사용된다.
- **장점:** 데이터를 직관적으로 비교할 수 있으며, 다양한 스타일(수직, 수평)을 사용할 수 있다.
- **단점:** 많은 범주가 있을 경우 가독성이 떨어질 수 있다.

2. 데이터 시각화 그래프 종류

성별 생존율을 막대그래프로 나타내줘.



2. 데이터 시각화 그래프 종류

티켓 등급별 생존율 막대그래프로 나타내줘.

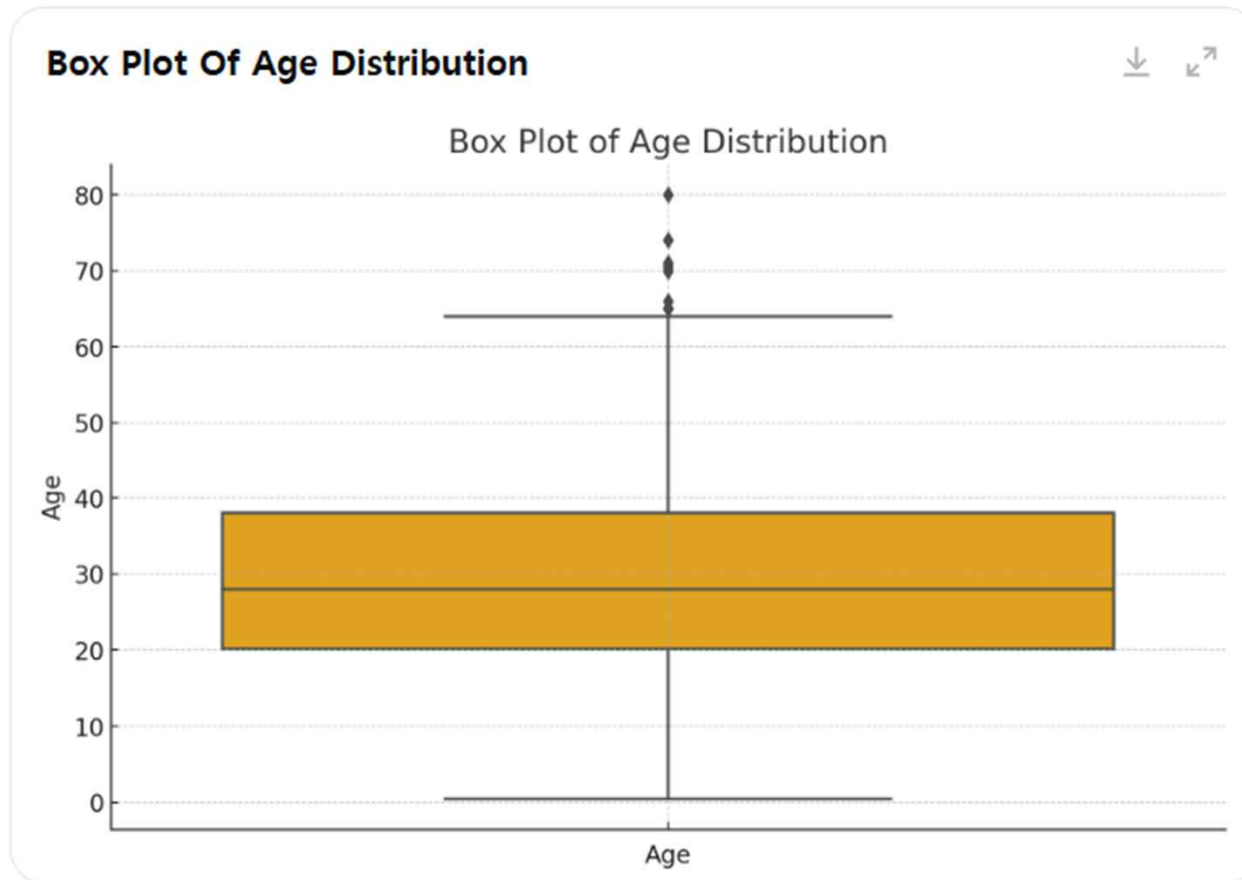


4. 박스 플롯 (Box Plot)

- **특징:** 데이터의 다섯 가지 요약(최소값, 1사분위수, 중앙값, 3사분위수, 최대값)과 이상치를 나타낸다.
- **용도:** 데이터 분포의 특성(중앙값, 사분위 범위)을 파악하고, 이상치를 식별할 때 유용하다. 예를 들어, 각 학급별 시험 점수의 분포를 비교할 때 사용된다.
- **장점:** 이상치를 쉽게 식별할 수 있고, 여러 그룹 간의 분포를 비교할 수 있다.
- **단점:** 데이터의 상세한 분포를 제공하지 않으므로, 데이터의 자세한 패턴을 알기 어렵다.

2. 데이터 시각화 그래프 종류

연령 분포를 박스플롯으로 그려줘.



2. 데이터 시각화 그래프 종류

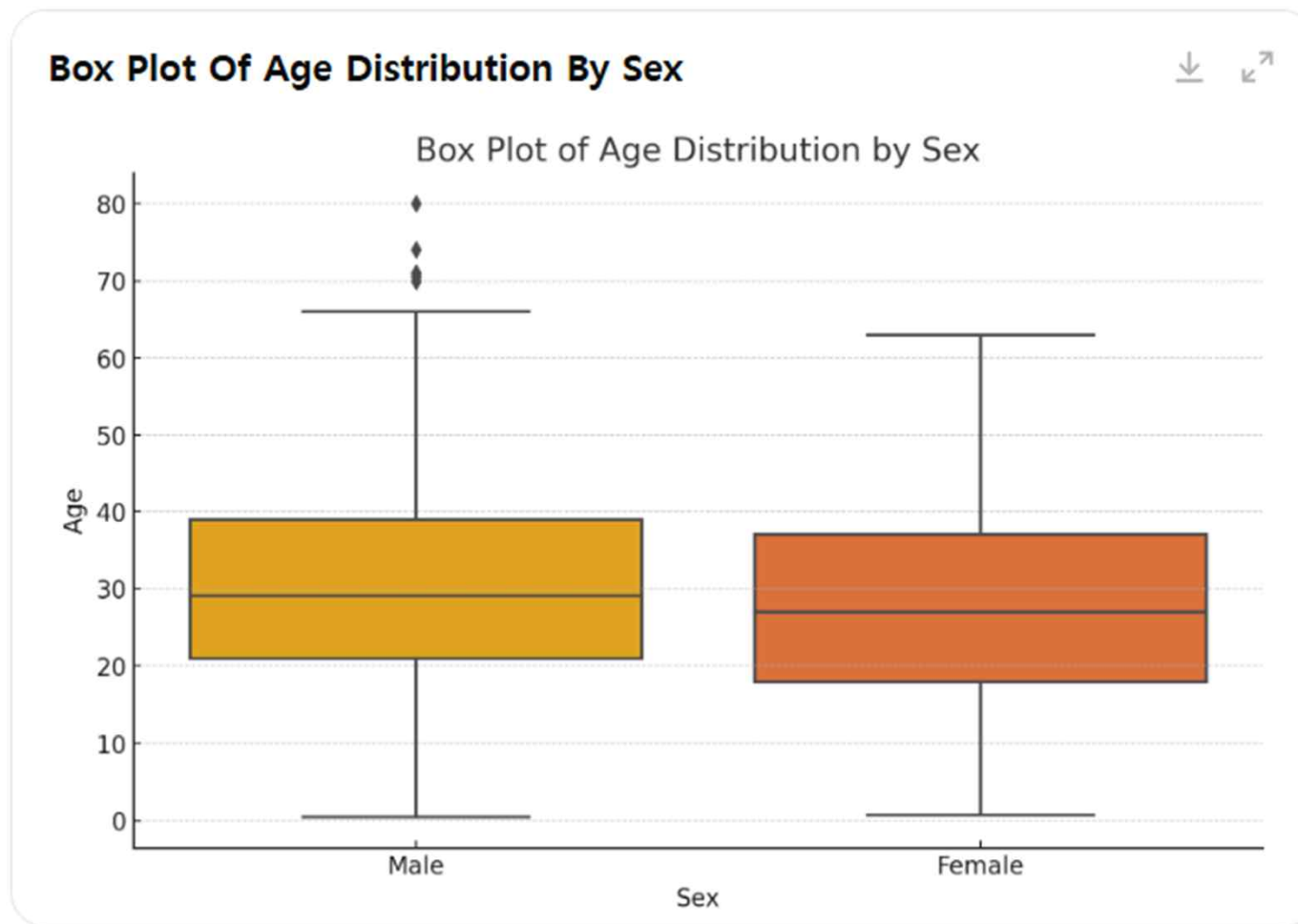
위의 박스플롯은 타이타닉 승객들의 연령 분포를 나타냅니다. 박스플롯을 통해 다음과 같은 정보를 얻을 수 있습니다:

- 중앙값(Median): 박스의 중앙에 있는 선은 승객들의 연령 중앙값을 나타냅니다.
- 사분위 범위(IQR): 박스의 위쪽과 아래쪽 경계는 각각 3사분위수(Q3)와 1사분위수(Q1)를 나타내며, IQR은 $Q3 - Q1$ 로 정의됩니다.
- 이상치(Outliers): 박스 위와 아래의 수염(whisker) 밖에 위치한 점들은 이상치로 간주됩니다.

박스플롯을 통해 대부분의 승객이 20대에서 40대 사이에 분포하고 있음을 알 수 있습니다. 또한, 몇몇 이상치로 간주되는 연령대가 있으며, 이는 매우 어린 승객이나 고령의 승객을 나타낼 수 있습니다. [↗]

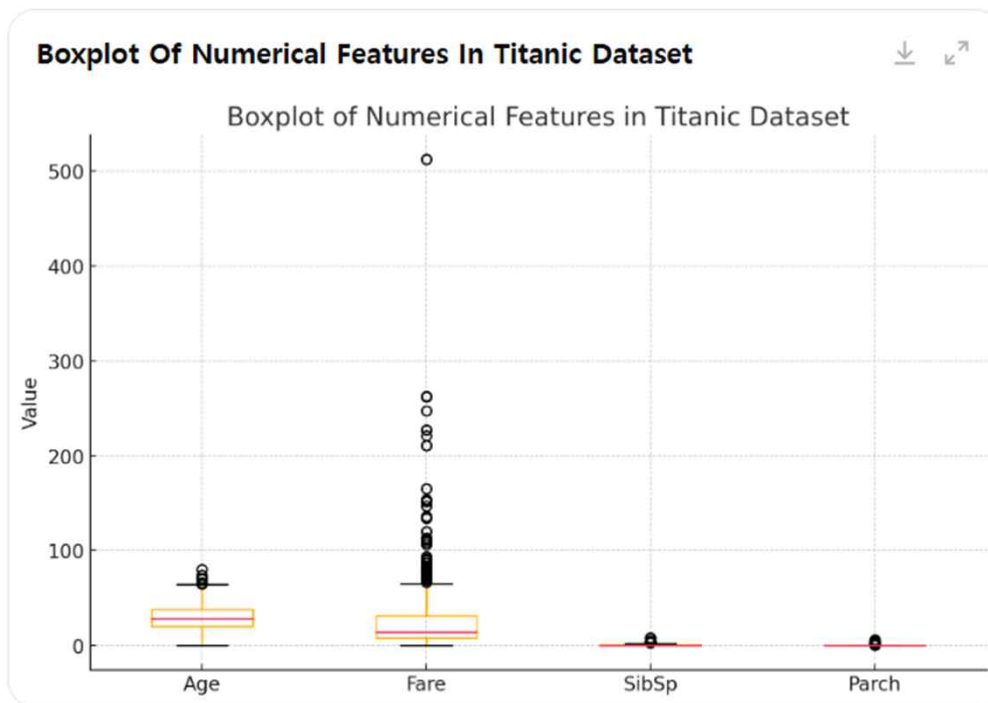
2. 데이터 시각화 그래프 종류

연령 분포와 성별 분포를 함께 박스 플롯으로 그려줘.



2. 데이터 시각화 그래프 종류

전체 분포를 박스 플롯으로 함께 그려줘.



Here is the boxplot displaying the distribution of numerical features (Age, Fare, SibSp, Parch) in the Titanic dataset. If you need further analysis or any other specific visualization, please let me know! [?]



숫자형만 출력함

5. 파이차트 (Pie Chart)

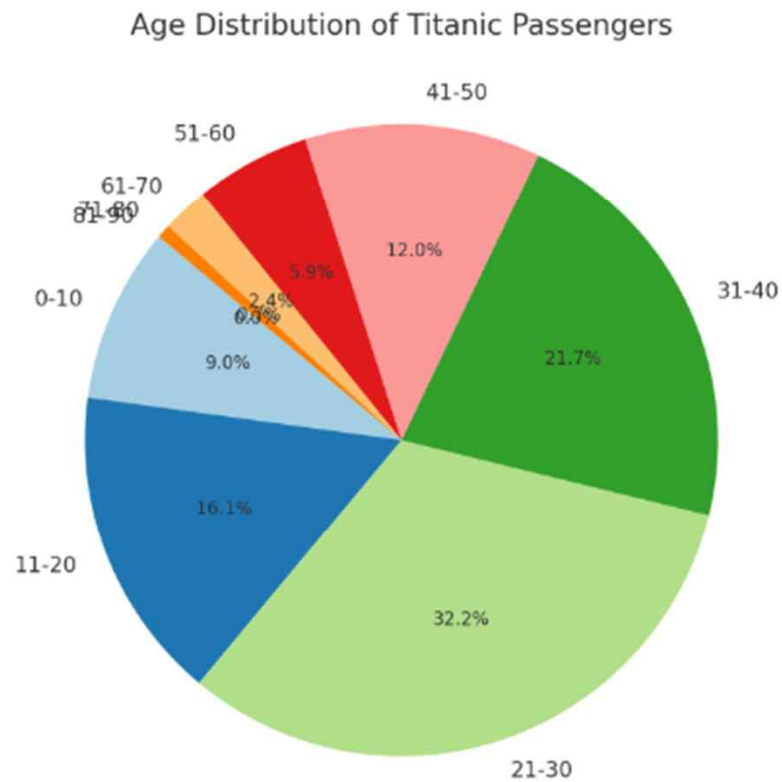
- **특징:** 전체에 대한 각 부분의 비율을 나타낸다.
- **용도:** 구성 요소들의 비율을 비교할 때 유용하다. 예를 들어, 시장 점유율, 예산 분포 등을 나타낼 때 사용된다.
- **장점:** 데이터를 직관적으로 이해할 수 있다.
- **단점:** 많은 조각이 있거나 비율 차이가 적을 경우 해석이 어려울 수 있다.

2. 데이터 시각화 그래프 종류

연령별 분포를 파이차트로 그려줘.



Age Distribution Of Titanic Passengers

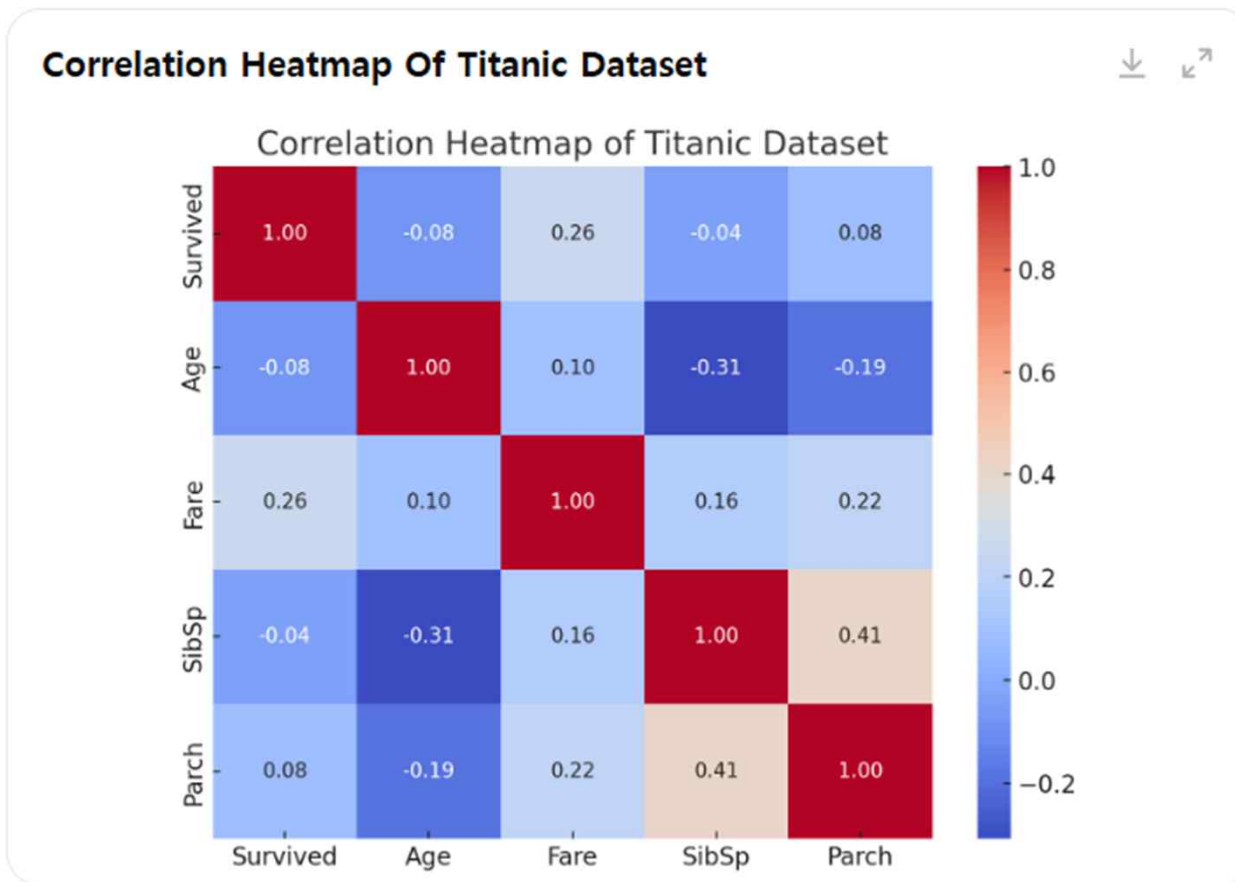


6. 히트맵 (Heatmap)

- **특징:** 값의 크기를 색상으로 표현하여 시각적으로 나타낸다.
- **용도:** 데이터 값의 강도나 빈도를 시각적으로 비교할 때 유용하다. 예를 들어, 웹사이트의 사용자 클릭 패턴, 상관 관계 매트릭스 등을 분석할 때 사용된다.
- **장점:** 복잡한 데이터셋의 패턴을 쉽게 파악할 수 있다.
- **단점:** 색상에 의존하기 때문에 색맹인 사람들에게는 해석이 어려울 수 있다.

2. 데이터 시각화 그래프 종류

각 변수와 Survived와의 상관 관계를 히트맵으로 그려줘.

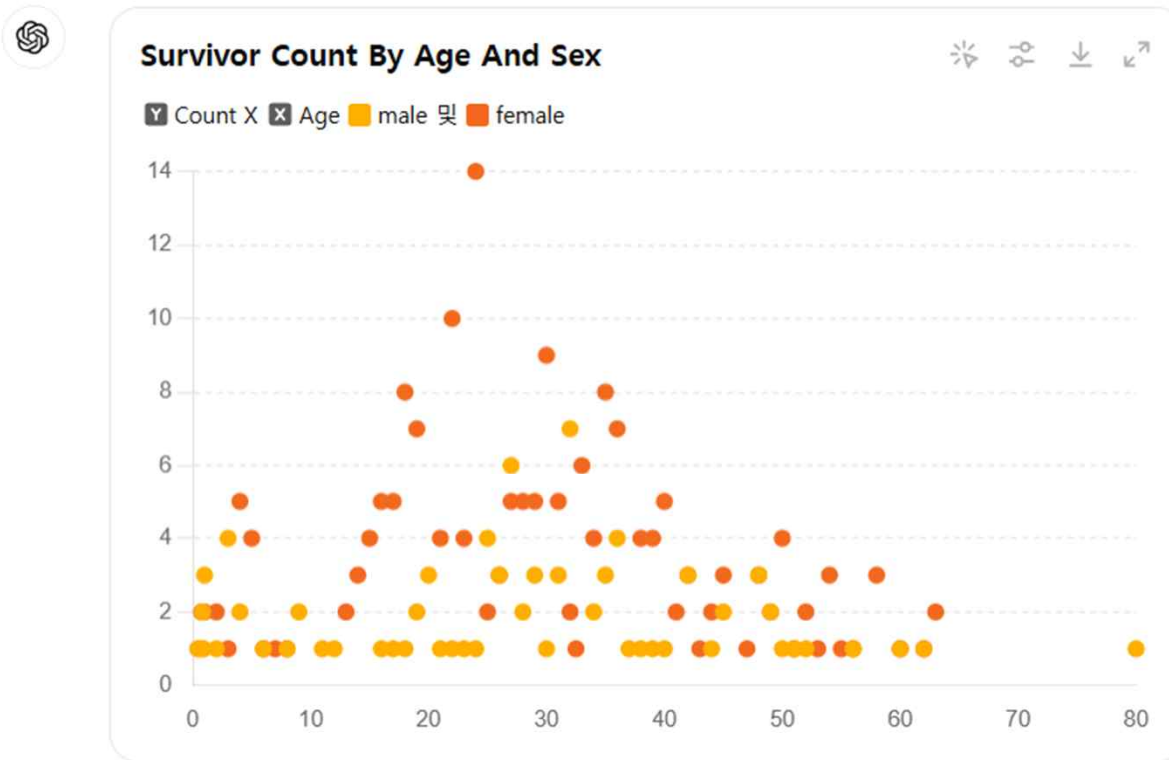


7. 버블 차트 (Bubble Chart)

- **특징:** 산점도에 버블의 크기를 추가하여 세 번째 변수를 표현한다.
- **용도:** 세 변수 간의 관계를 시각화할 때 유용하다. 예를 들어, 각 도시의 인구, 소득, 인프라 수준을 비교할 때 사용된다.
- **장점:** 여러 변수를 동시에 표현할 수 있다.
- **단점:** 버블이 겹치면 해석이 어려울 수 있다.

2. 데이터 시각화 그래프 종류

연령과 성별에 따른 생존자수를(사망자수는 제외) 버블차트로 그려줘.



Here is the bubble chart displaying the count of survivors by age and sex. The size of the bubbles represents the number of survivors in each category. If you need further analysis or any other adjustments, feel free to let me know! [?]

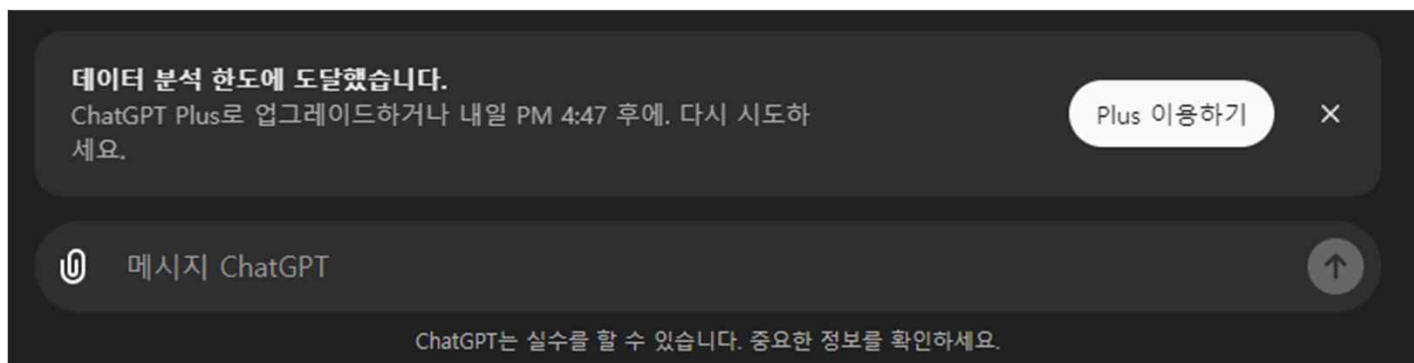
2. 데이터 시각화 그래프 종류

데이터 분포를 시각화해줘. [ChatGPT 무료버전 사용]



2. 데이터 시각화 그래프 종류

데이터 분석을 위한 시각화 출력해줘. [ChatGPT 무료버전 사용]
(1일 사용한도 초과)



2. 데이터 시각화 그래프 종류

[참고] **GPT-4 사용 한도 초과시 사용 중지됨** : 3시간에 40개 메시지 이내



현재 GPT-4 사용 한도에 도달했습니다. 지금 기본 모델을 계속 사용하거나 다시 시도하실 수 있습니다 after 12:05 AM. [자세히 알아보기](#)

기본 모델 사용하기

< 2/2 >

바로 3시간 마다 50개의 질문만을 사용할 수 있다는 것입니다. 이는 ChatGPT4의 모든 사용자들에게 GPT-4를 공정하게 기회를 주기 위한 것이라고 합니다. 만약 3시간에 50개의 메시지를 초과하면, GPT-4를 사용할 수 없게 되고, GPT-3.5라는 무료 버전의 모델로 자동으로 전환됩니다. GPT-3.5는 GPT-4보다는 답변이 단조로워 정확성이 떨어지며 복잡한 작업에는 한계가 있습니다.

<https://seasonf.tistory.com/74>

감사합니다