

5. 공공데이터(Titanic) 분석

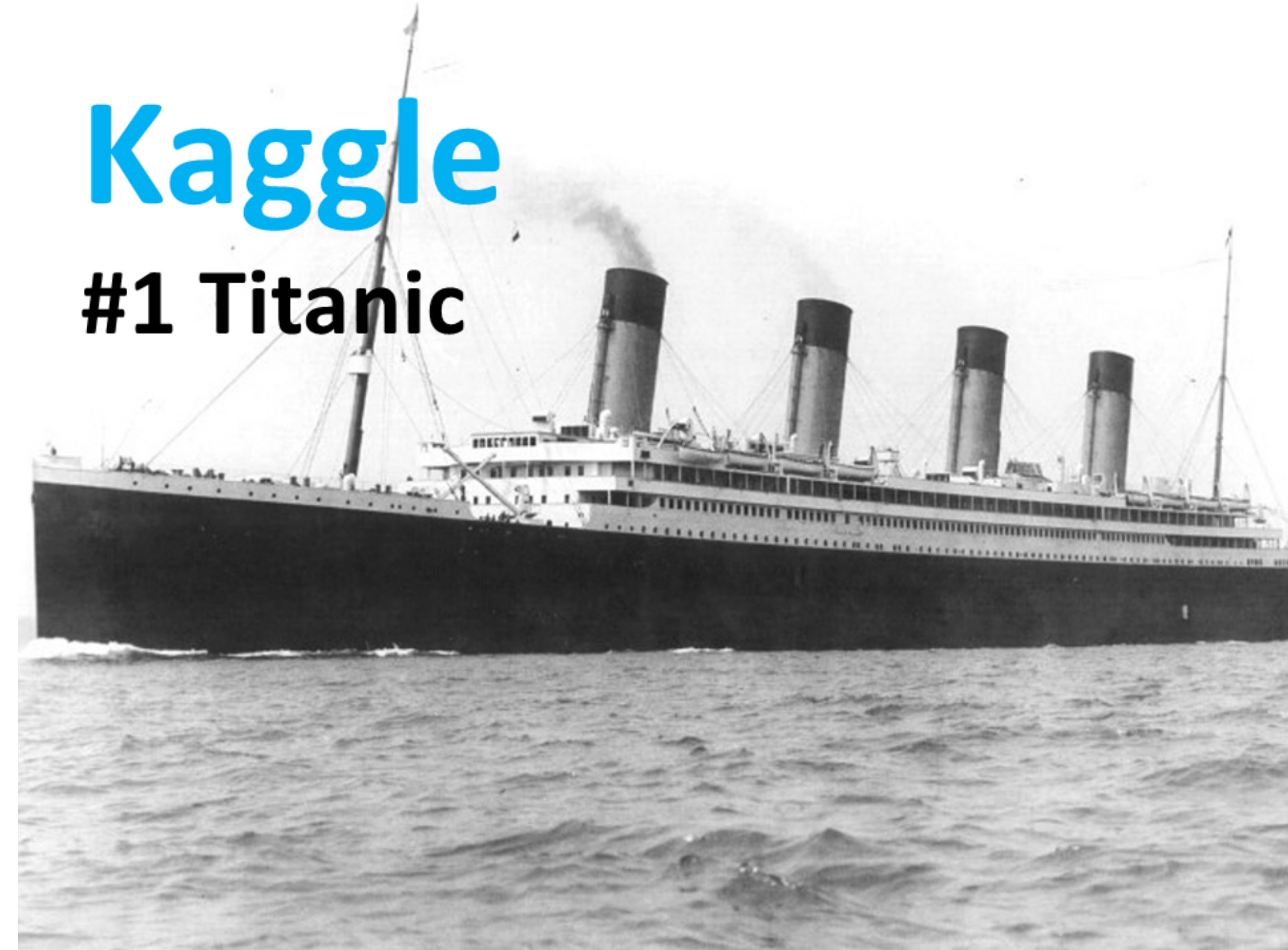
EDA 실습 (Kaggle)

8월 25일

학습목표

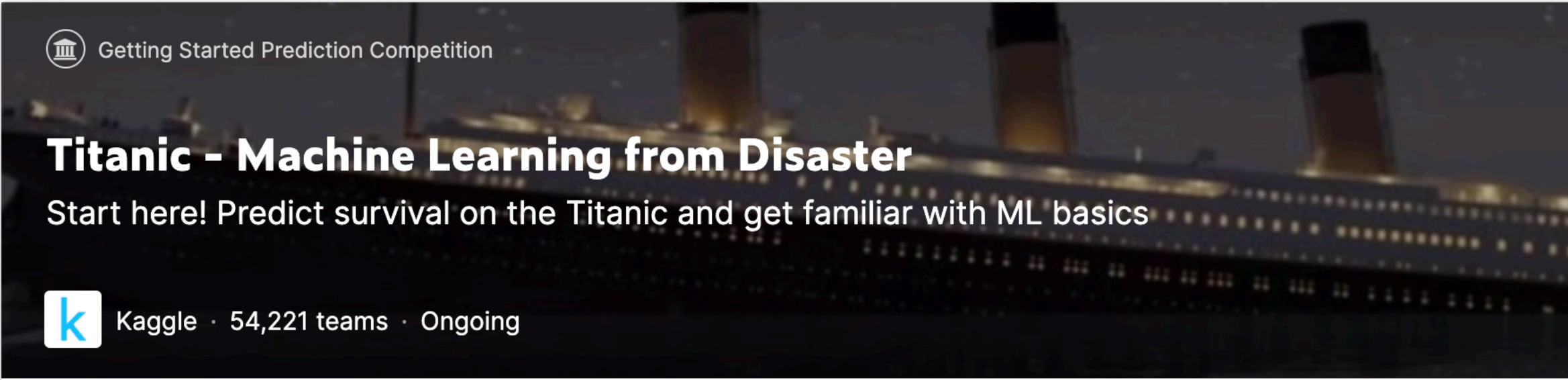
수행결과

- 타이타닉 승객 파일에서 여러 가지 정보를 추출해 본다
- 승객 중에서 최고령자가 누구인가 ?
- 타이타닉 호 침몰 사건 당시의 사망자와 생존자를 구분하는 요인 분석
- 어떤 승객이 생존 가능성이 높은가 ? 남자 (여자), 1등석 등



Pandas


Titanic 데이터셋



Getting Started Prediction Competition

Titanic - Machine Learning from Disaster

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

 Kaggle · 54,221 teams · Ongoing

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Join Competition](#) [...](#)

Data Description

Overview

The data has been split into two groups:

- training set (train.csv)
- test set (test.csv)

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

데이터셋 개요

타이타닉 데이터셋

- 타이타닉 탑승자에 대한 데이터셋 train.csv를 다운로드 받음

[Overview](#)[Data](#)[Code](#)[Discussion](#)[Leaderboard](#)[Rules](#)[Team](#)[My Submissions](#)[Submit Predictions](#)[...](#)

Data Explorer

90.9 KB

- gender_submission.csv
- test.csv
- train.csv

< train.csv (59.76 KB)

Download

Fullscreen

DetailCompactColumn

10 of 12 columns

About this file

contains data

PassengerId

Survived

Pclass

Name

1891

01

13

891 unique values

1

0

3

Braund, Mr. O'Hara

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch
1	0	3	Braund, Mr. O'Hara	male	22	1	0
2	1	1	Allen, Mr. William	male	35	0	0
3	1	3	Harris, Mrs. Charles	female	26	0	0
4	1	1	Scott, Mrs. James	female	35	1	0
5	0	3	Harris, Mr. James	male	35	0	0

데이터셋 개요

데이터 딕셔너리

- PassengerId: 승객의 ID
- Survived: 생존 여부
- Pclass: 탑승 등급을 나타낸다. 클래스 1, 클래스 2, 클래스 3로 구성
- Name: 승객의 이름
- Sex: 승객의 성별
- Age: 승객의 나이
- SibSp: 승객에게 형제 자매와 배우자가 있음을 표현
- Parch: 승객이 혼자인지 또는 가족이 있는지 여부를 표현
- Ticket: 승객의 티켓 번호
- Fare: 운임
- Cabin : 승객의 선실
- Embarked: 탑승한 지역

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

데이터셋 개요

- 승객 891명에 대한 데이터를 포함
- 각 행은 탑승자 1인의 정보를 나타냄 (생존여부 포함)
 - SibSp 형제 자매, 배우자 유무
 - Parch 혼자인지 또는 가족이 있는지 유무
 - Cabin 승객의 선실

분석파이프라인

데이터준비

- 판다스 패키지 import 하며 pd 별명 지정하기

```
import pandas as pd
```

- Titanic 데이터셋 읽고 데이터프레임 df 구성하기

```
titanic_df = pd.read_csv(url)
```

- Titanic 데이터셋 확인하기

```
titanic_df.head()
```


분석파이프라인

데이터보기

- 데이터프레임의 기본정보 출력

titanic_df.info()

- 데이터프레임 기초통계확인

titanic_df.describe()

	PassengerId	Survived	Pclass	Age	SibSp	Parch
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000

```
[ ] titanic_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
#   Column          Non-Null Count  Dtype
---  -
0   PassengerId      891 non-null    int64
1   Survived         891 non-null    int64
2   Pclass           891 non-null    int64
3   Name             891 non-null    object
4   Sex              891 non-null    object
5   Age              714 non-null    float64
6   SibSp            891 non-null    int64
7   Parch            891 non-null    int64
8   Ticket           891 non-null    object
9   Fare             891 non-null    float64
10  Cabin            204 non-null    object
11  Embarked         889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```


분석파이프라인

자료 특징 확인

탐색 대상 변수	두 그룹 간의 분포 혹은 평균의 차이가 있는가?
pclass	○
age	X
sibsp, parch	△
fare	○
sex	○
embarked	△

분석파이프라인

데이터보기

- 데이터 유형 확인

```
titanic_df.dtypes
```

- 데이터 행 인덱스 확인

```
titanic_df.loc[0]
```

- 데이터 앞 자료 추출

- 데이터 뒤 자료 추출

- 데이터프레임 인덱스 보기

- 행, 열 구조 보기

분석파이프라인

데이터보기

- 요금 기준 오름차순 정렬

```
titanic_df.sort_values(by=['Fare'], axis=0)
```

- 요금 기준 내림차순 정렬

```
titanic_df.sort_values(by=['Fare'], axis=0, ascending=False)
```

- 열이름을 알파벳 순으로 정렬하기

```
titanic_df.sort_index(axis=1)
```

분석파이프라인

자료에서 특정 정보 얻기

- 승객나이 추출하기
- 타이타닉 탑승객 중에서 최고령 나이 확인하기
- 타이타닉 승객데이터 기본 통계 확인하기(생존율 확인)
- 타이타닉 탑승객 평균 나이를 계산하기

분석파이프라인

결측치 다루기

- 자료에서 결측치를 확인한다

```
titanic_df.count() # 데이터 개수 확인  
titanic_df.isnull().sum() # 결측치 확인
```

- 확인된 결측치 대체/제거를 통해 데이터 클리닝 수행
 - Age는 중간값으로 대체
 - Cabin은 해당 열을 제거
 - Embarked는 최대빈도값으로 대체

(실습) 데이터준비

- 승객 ID 인 PassengerID 를 인덱스로 지정하여 자료를 읽어 데이터프레임을 구성한다

```
titanic_df = pd.read_csv(url, index_col='PassengerId')  
titanic_df.head()
```


자료에서 특정 정보 얻기

- 타이타닉 탑승객의 이름, 나이, 성별 정보 얻기 (해당 자료를 df1 프레임워크로 구성)
- 20세 미만의 승객언어 필터링하기 (조건에 맞는 행을 below_20 프레임워크로 구성)
- 1등석, 2등석에 탑승 승객 출력하기
- 1등석, 2등석에 탑승 승객 이름 출력하기 (df.loc [조건, 열레이블])

자료 통계

- 타이타닉 승객의 평균 연령 구하기
- 타이타닉 승객 연령 (요금)의 중간값 구하기

데이터 집계

데이터 그룹 분석(group analysis)

- 특정조건에 맞는 데이터가 하나 이상 데이터 그룹을 이루는 경우 해당 집단 특성을 보여주기 위해 사용
- 그룹분석 과정
 - 범주형 필드를 기준으로 정하여 그룹으로 데이터 분할(splitting)
 - 그룹에 평균, 합 등의 독립적인 함수 적용(apply)
 - 결과물을 하나의 데이터 구조로 결합(combining)

데이터 집계

데이터 그룹 분석(group analysis)

- 선실별로 자료를 그룹핑한 후 평균연령 확인해 본다

```
titanic_df.groupby('Pclass').mean()
```

- 승객의 선실 등급의 성별에 따른 평균 요금 구하기

```
titanic_df.groupby(['Pclass', 'Sex'], as_index=False).mean()
```

- 선실의 성별에 따른 평균연령 구하기

```
part_df = titanic_df[['Pclass', 'Sex', 'Age']]  
part_df.groupby(['Pclass', 'Sex']).mean()
```

데이터 집계

데이터 그룹 분석(group analysis)

- 선실의 남녀별 최고령 나이 구하기

```
part_df.groupby(['Pclass', 'Sex'], as_index=False).max()
```