

# 5일차. 캐글 데이터 분석 실습1

심선영 교수, 이주민 교수

# 강의 목표

- ❖ 파이썬 기초 복습 및 확장하기를 통해 기본기를 다진다.
- ❖ 캐글 사이트를 가입하고 및 타이타닉 대회를 이해한다
- ❖ 타이타닉 데이터를 이용하여 탐색적 데이터 분석을 한다.

# 강의 스케줄

목차	활동
데이터 처리 실습	- 파이썬 문법
	- 데이터 처리
	- 시각화
캐글 대회 준비	캐글 가입하기 & Titanic 대회 설명
캐글 데이터 분석	Titanic 탐색적 데이터 분석 - Day5_4 titanicEDA.ipynb
Wrap-Up	

Review

캐글 실습

# 캐글 가입하기

The image shows the Kaggle homepage. At the top, there's a navigation bar with 'kaggle', a search bar, and links for 'Competitions', 'Datasets', 'Notebooks', 'Discussion', 'Courses', 'Sign in', and 'Register'. The main content area features a large banner with the text 'Start with more than a blinking cursor' and 'Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access free GPUs and a huge repository of community published data & code.' Below this, there are two prominent buttons: 'REGISTER WITH GOOGLE' and 'Register with Email'. Red annotations highlight these buttons with the text '구글 계정으로 시작하기' (Start with Google account) and '이메일로 시작하기' (Start with email).

The image shows the Kaggle sign-in and registration screen. It has a 'kaggle' logo at the top. Below the logo, there are two tabs: 'Sign In' and 'Register'. The 'Register' tab is selected. Under the 'Register' tab, there are two buttons: 'Register with Google' and 'Register with your email'. At the bottom, there's a link that says 'Have an account? Sign in.'

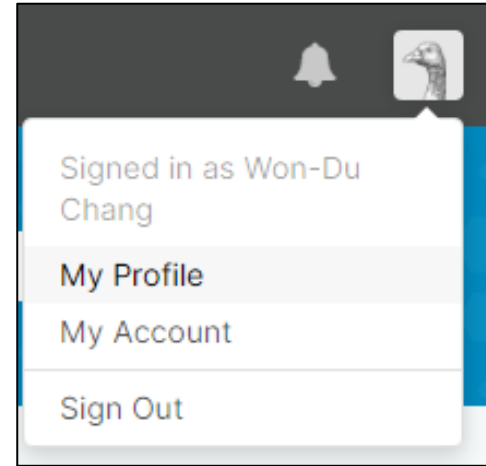
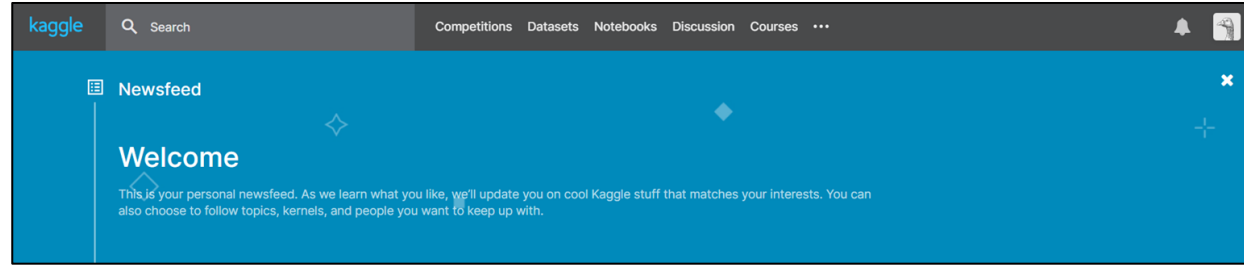
The image shows the Kaggle registration form. It has a 'kaggle' logo at the top. Below the logo, there's a 'back' link. The form is titled 'Register'. It contains three input fields: 'Email address', 'Password (min 7 chars)', and 'Full name (displayed)'. Red annotations highlight these fields with the text '이메일 주소' (Email address), '비밀번호' (Password), and '영문이름' (English name). Below the input fields, there are two checkboxes: 'I am not a robot.' and 'Subscribe to newsletter'. At the bottom, there are 'Cancel' and 'Next' buttons.

The image shows the Kaggle Privacy and Terms screen. It has a 'kaggle' logo at the top. Below the logo, there's a 'Privacy and Terms' section. It contains a scrollable area with text about the terms of use. At the bottom, there are 'Cancel' and 'I agree' buttons. A red annotation highlights the 'I agree' button with the text '동의' (Agree).

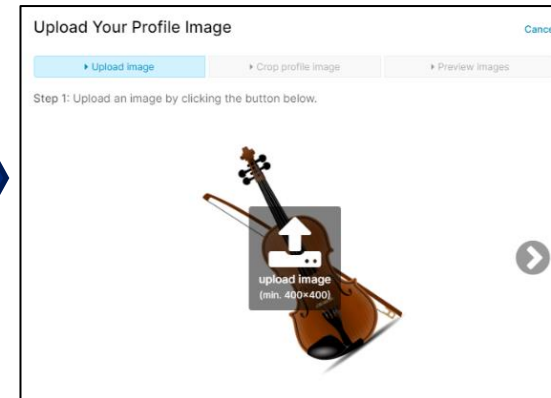
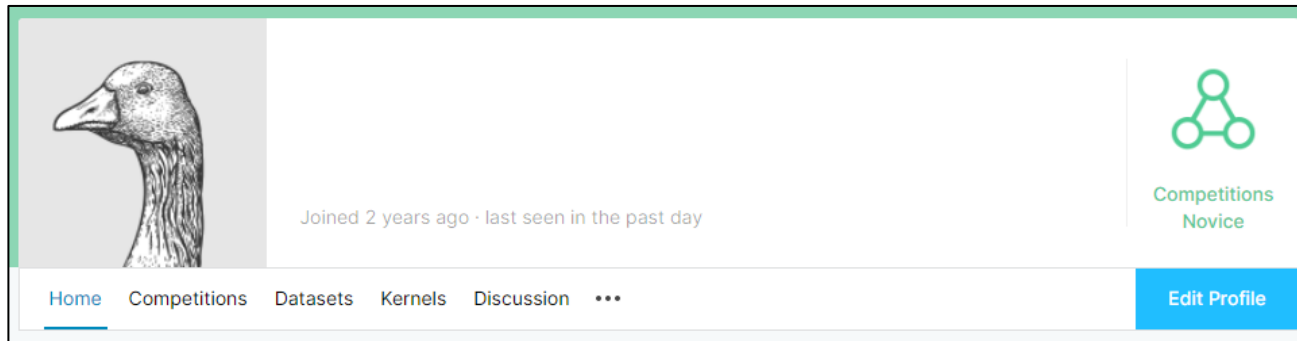
The image shows the Kaggle 'Verify your email' screen. It has a 'kaggle' logo at the top. Below the logo, there's a 'Verify your email' section. It contains a text input field for a 'Six-digit code'. Red annotations highlight this field with the text '인증번호 입력' (Enter verification number). Below the input field, there are 'Resend email' and 'Next' buttons. A red annotation highlights the 'Next' button with the text '다음으로 진행' (Proceed to next).

캐글입문

# 캐글 가입하기



- 가입완료 후
- 프로필 보기

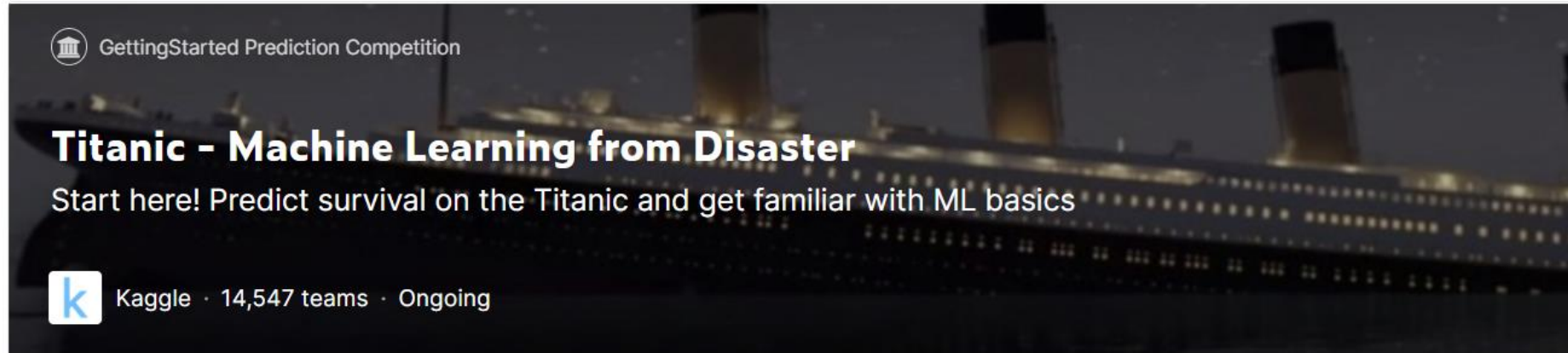


가입한 것 확인!

# Titanic 대회 이해하기




# Titanic Competition 개요



GettingStarted Prediction Competition

## Titanic - Machine Learning from Disaster

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

 Kaggle · 14,547 teams · Ongoing


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

Overview

Description

Evaluation

Frequently Asked Questions

 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.

The competition is simple: use machine learning to create a model that predicts which passengers survived the Titanic shipwreck.

# Titanic Competition 개요

## ❖ 배경지식

- RMS 타이타닉호는 1912년 4월 10일 영국의 사우스햄프턴을 떠나 미국의 뉴욕으로 향하던 첫 항해 중에 4월 15일 빙산과 충돌하여 침몰하였다. 타이타닉호의 침몰로 1,514명이 사망하였다.

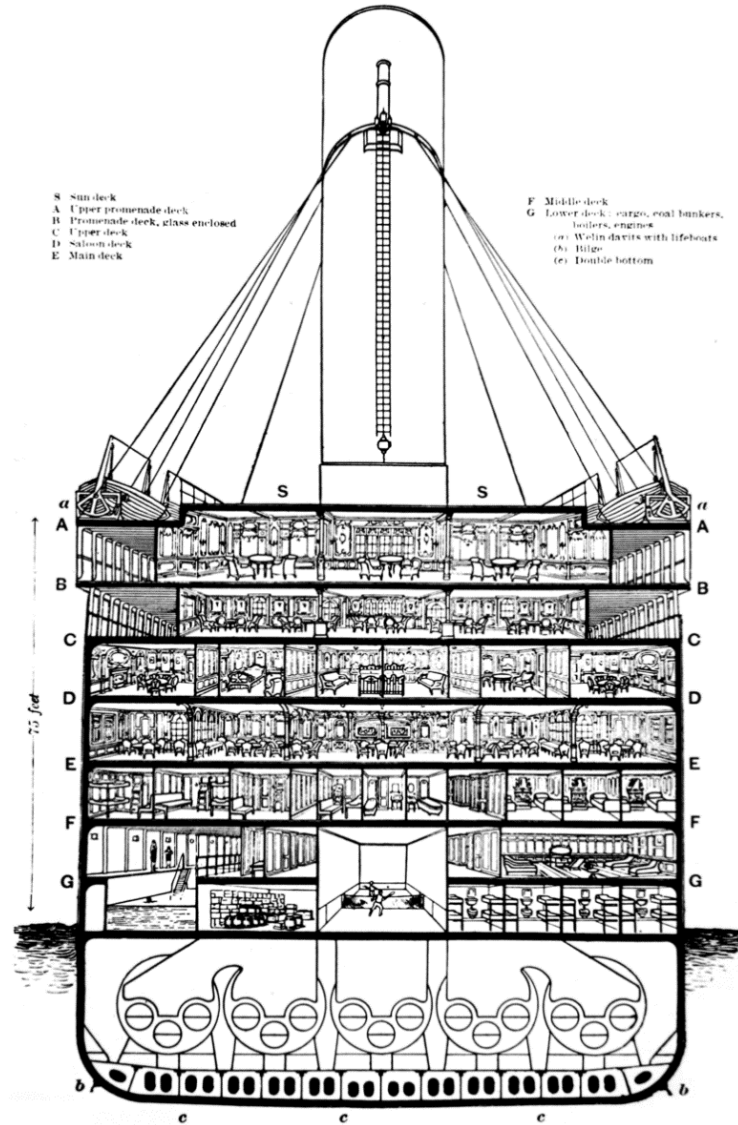
## ❖ 목표

- 생존유무가 포함된 탑승객의 학습 데이터 정보를 분석하고 그 정보를 바탕으로 기계학습을 진행하여 테스트 데이터의 탑승객 생존유무를 예측

# Titanic-Machine Learning from Disaster Competition



영화(타이타닉) 포스터



출처: 위키피디아: 타이타닉

## 타이타닉 호

- 1912년 침몰한 대형 증기선
- 가라앉지 않는 배로 불렸으나 첫 항해에서 침몰
- 2천명 이상의 승객 중 1,500여명이 숨진 비극적 사건
- 1997년에 영화화됨

## 생존자 통계 데이터

- 나이, 성별, 탑승권 등급 등에 따라 생존률이 다름
- 캐글에서 이 데이터셋을 사용하여 생존여부를 예측하는 대회 open



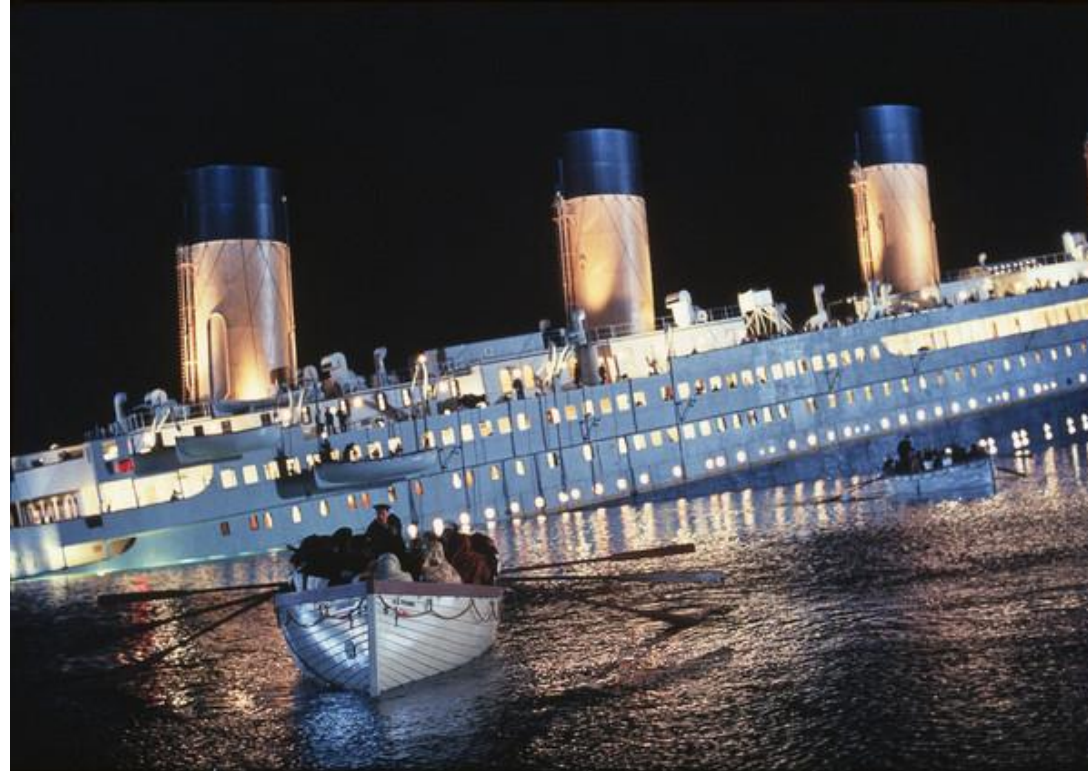
# Titanic-Machine Learning from Disaster Competition

## ❖ 타이타닉 사건 이해하기



# Titanic-Machine Learning from Disaster Competition

❖ 타이타닉 사건 이해하기





# 데이터 분석 프로세스별 실습

## ❖ 타이타닉 사건 이해하기

Embarked	탑승지 (C: 세르부르, Q: 퀸즈타운, S: 사우스햄프턴)
----------	-----------------------------------

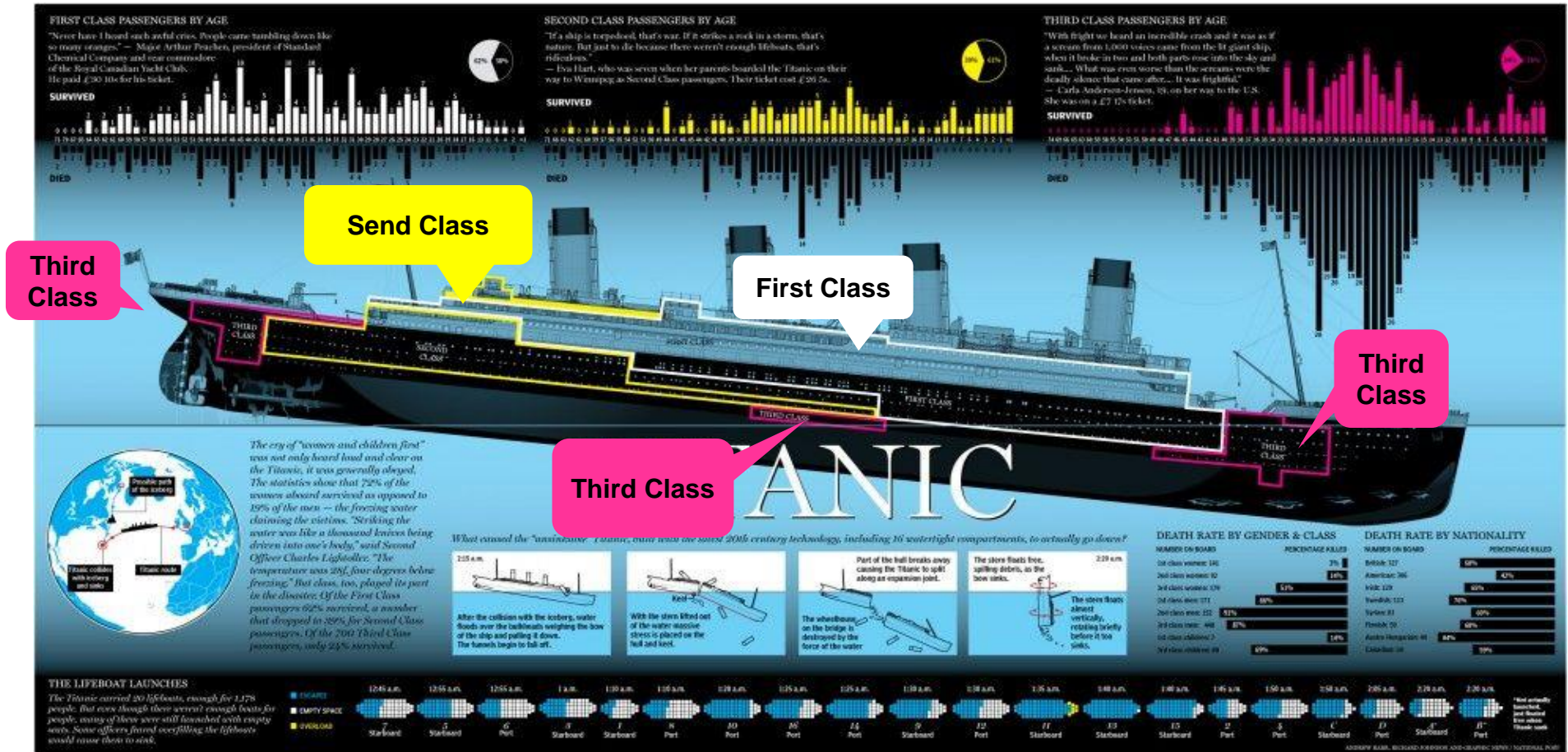


**1912년 4월 10일** 처녀항해 (처음 출항한다는 의미) **프랑스 출발**  
**1912년 4월 17일** 미국 뉴욕에 닿을 예정

# 데이터 분석 프로세스별 실습

## ❖ 타이타닉 사건 이해하기

Pclass	탑승권 종류 (1: 1등석, 2: 2등석, 3: 3등석)
--------	---------------------------------



# Titanic Competition 개요

## ❖ Data Dictionary

- Survived (생존유무): 0 = 사망, 1= 생존
- Name: 탑승객 이름
- pclass: Ticket class (티켓클래스) 1= 1st, 2=2nd, 3=3rd
- Sex: 성별 male, female
- Age: 나이(세)
- sibsp: # of siblings/spouses aboard the Titanic(함께 탑승한 형제자매, 배우자 수 총합)
- parch: # of parents/children aboard the Titanic(함께 탑승한 부모, 자녀 수 총합)
- ticket: Ticket Nubmer(티켓 넘버)
- cabin: Cabin Number(객실 넘버)
- embarked: Port of Embarkatation (탑승항구) C = Cherbourg, Q= Queenstown, S=Southampton



# 탐색적 데이터 분석

## ❖ EDA (Exploratory Data Analysis)

- 데이터의 종류 및 특징을 확인
- 데이터 간의 관계 파악, 가설을 수립하기 위한 기법
- 통계 및 시각화 기법을 사용
- 데이터의 전처리 과정을 포함
  - 결측치 처리, 이상치(outlier) 제거
  - 데이터 변환 (차원축소, 인코딩, 로그변환)
- EDA가 중요한 이유
  - 데이터 내의 패턴 및 비정상적 형태 등을 파악, 변수간의 관계 이해(가설) →모델 수립을 정교화

(!!)데이터를 탐색, 가공, 분석 하기 전에 먼저 알아야 할 것이 있음! → 데이터와 척도의 유형

Wrap-Up

# Day 5 Wrap-up 설문

❖ <https://forms.gle/1y7rMifoWdXYEg9bA>