# 探索性資料分析(EDA)_數據理解與重覆和遺失值處理

簡報閱讀 — 範例與作業 — 問題討論 — 學習心得(完成)

重要知識點 ›
什麼是 kaggle？ ›
鐵達尼號生存預測 ›

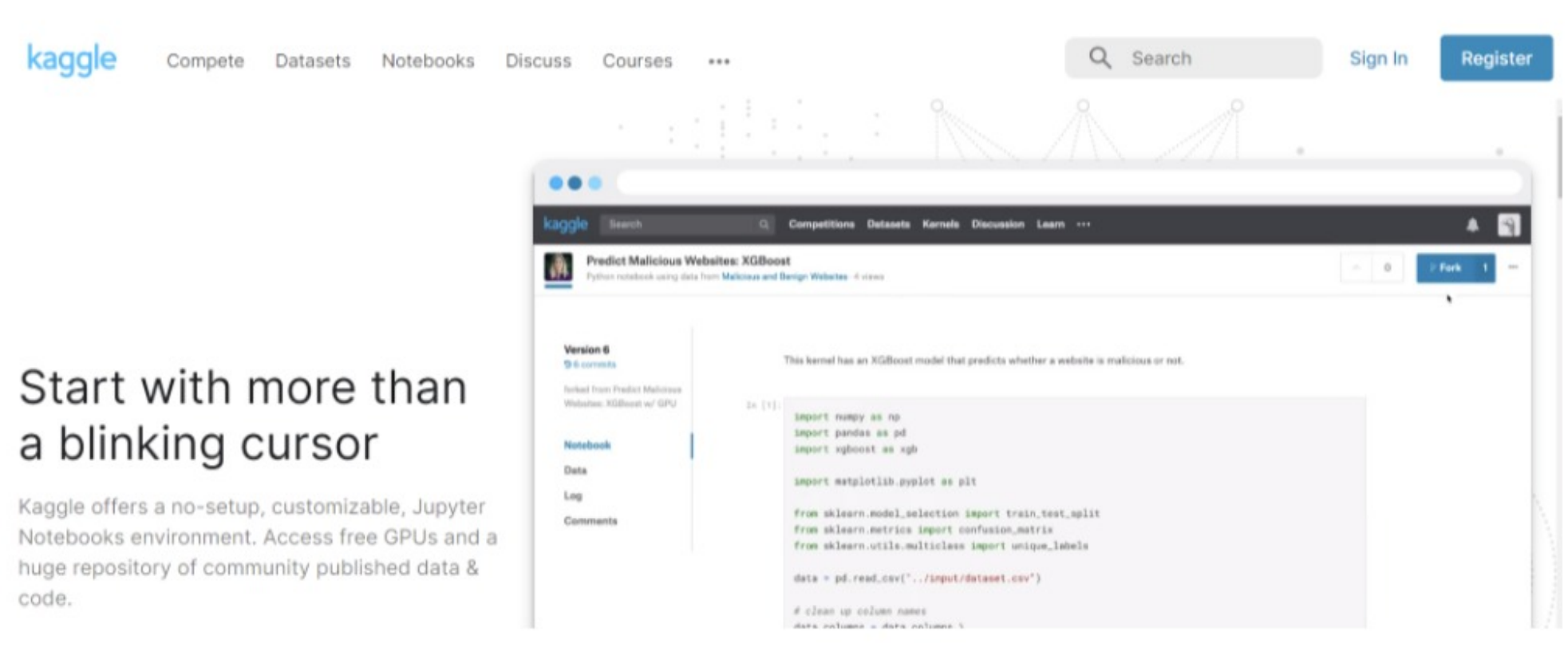
## 重要知識點

- 以 kaggle 鐵達尼資料集為例，理解資料科學數據議題
- 理解遺失值的特性
- 運用 python ，處理資料中重覆和遺失值處理

## 什麼是 kaggle？

網站：kaggle

Kaggle 早期因與各領域的公司單位合作舉辦數據分析競賽而出名；
競賽的舉辦由與其合作的公司或單位來定義想解決的問題並提供相關數據資料，然後開放給各路好手建立解決問題的預測模型。



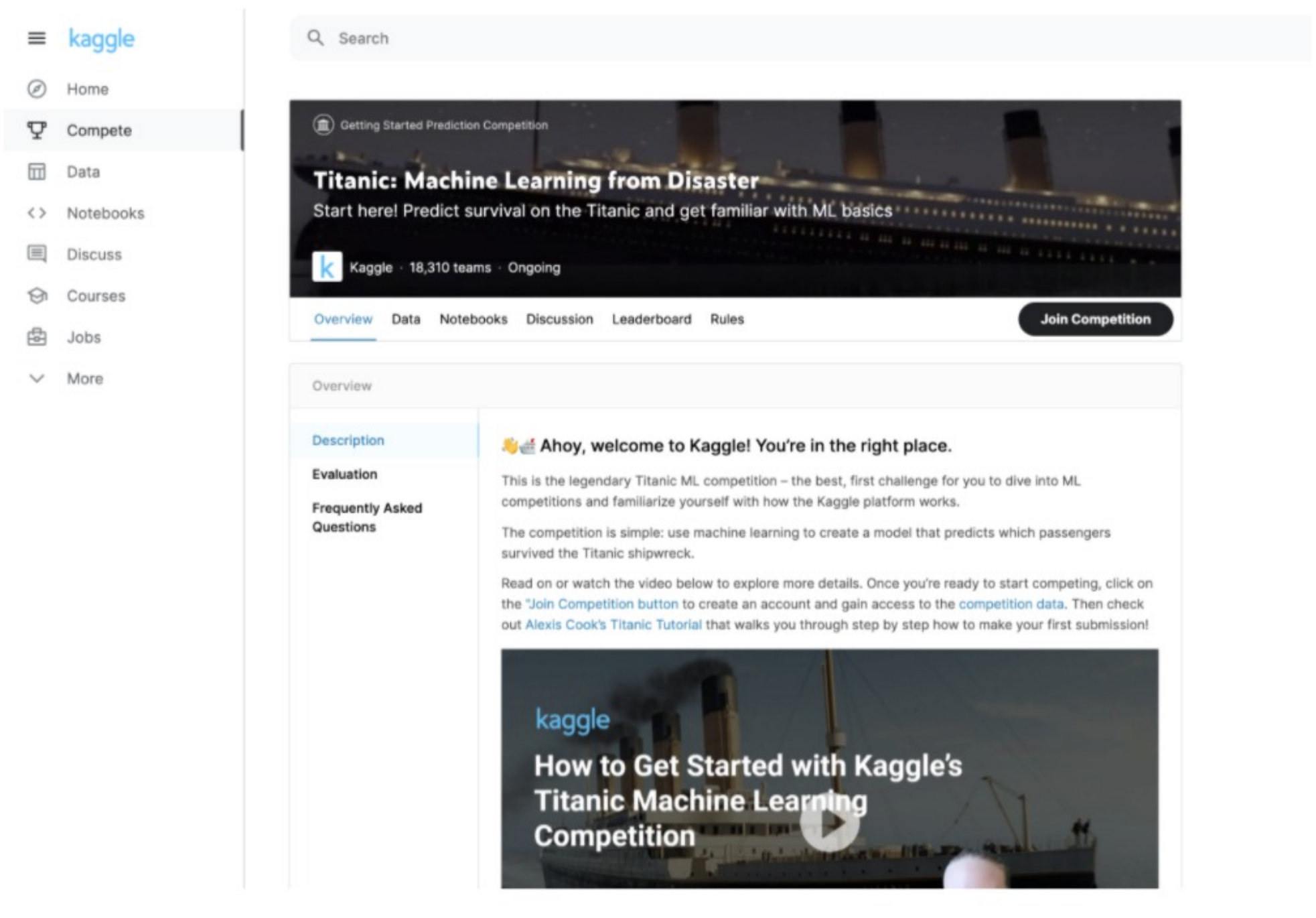
Start with more than a blinking cursor

Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access free GPUs and a huge repository of community published data & code.

## 鐵達尼號生存預測

預測怎樣的組合條件比較容易存活？

資料集連結：https://www.kaggle.com/c/titanic



Titanic: Machine Learning from Disaster

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

Kaggle · 18,310 teams · Ongoing

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.

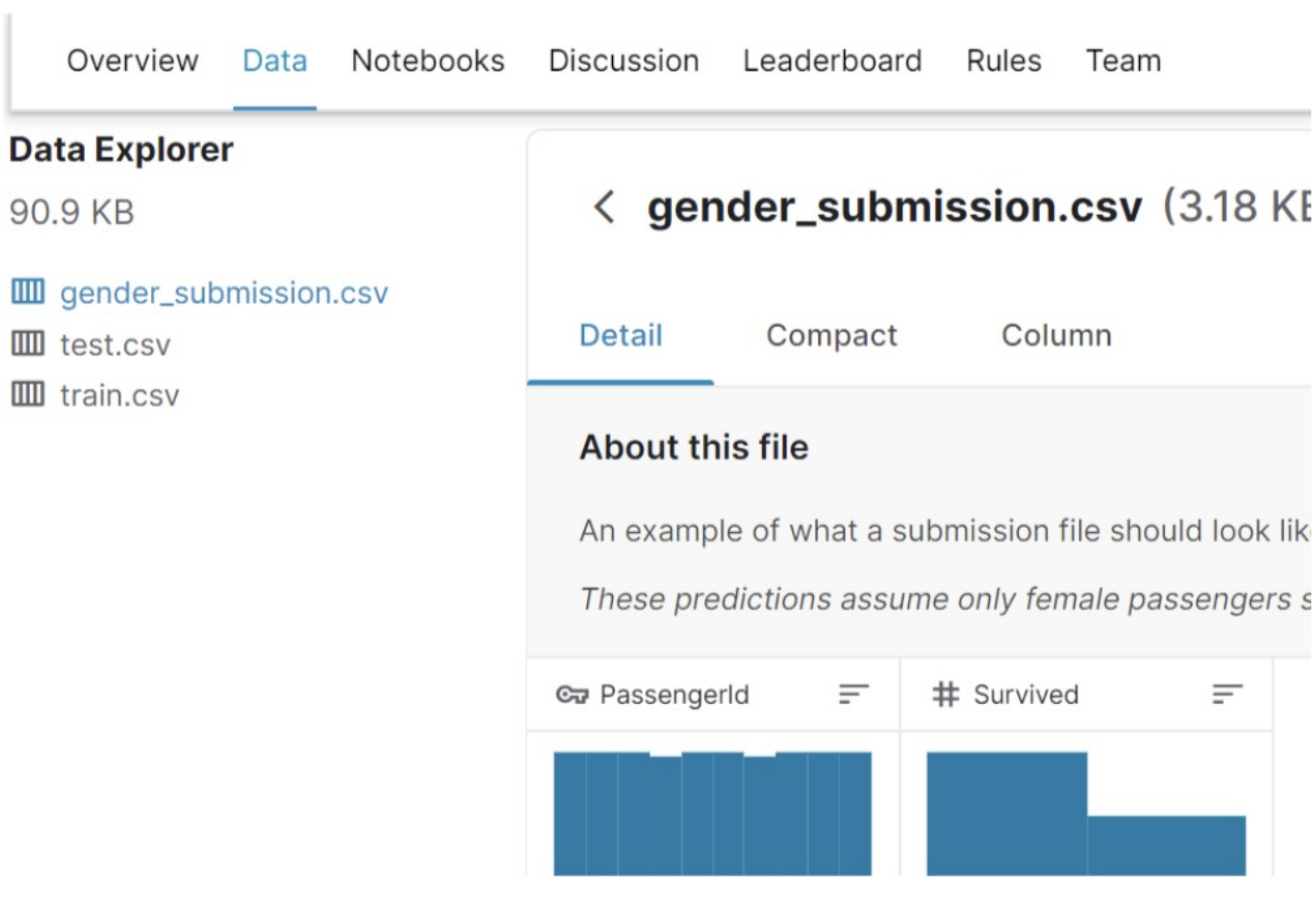
Read on or watch the video below to explore more details. Once you're ready to start competing, click on the "Join Competition button to create an account and gain access to the competition data. Then check out Alexis Cook's Titanic Tutorial that walks you through step by step how to make your first submission!

## 資料描述

- 鐵達尼號沈船事件，發生在 1912 年 4 月 15 日
- 船上共 2224 名乘客，共 1502 名死亡。
- 在過程中，發現某些族群容易存活下來。
- 目標：預測怎樣的組合條件比較容易存活？

## 資料架構

| 資料 | 意義 |
|---|---|
| train.cvs | 訓練資料集，訓練模型 |
| test.csv | 測試資料集，以此資料透過模型，產生預測值 |
| Gender_submission.csv | Kaggle 上傳的資料格式。 |



About this file

An example of what a submission file should look lik

These predictions assume only female passengers s

資料來源：kaggle titanic data

## 釐清每個資料欄位定義

| 變數(Variable) | 意義(Meaning) | 資料型態 |
|---|---|---|
| PassengerId | 乘客編號 | 數字 |
| Survived | 是否存活 (1:活/0:死) | 離散 |
| Pclass | 票務艙 (1:Upper,2:Middle,3:Lower) | 離散 |
| Name | 姓名 | 字串 |
| Sex | 性別 | 離散 |
| Age | 年齡 | 連續 |
| SibSp | 在船上兄弟姊妹配偶的人數<br>(定義家庭關係，非家庭關係不納入) | 連續 |
| Parch | 在船上父母和子女的人數<br>(定義家庭關係，非家庭關係不納入) | 連續 |
| Ticket | 船票號碼 | 字串 |
| Fare | 乘客票價 | 字串 |
| Cabin | 船艙號碼 | 離散 |
| Embarked | 登船港口 | 離散 |

## 載入鐵達尼資料，發現有遺失值

為什麼要處理遺失值？

- 有些程式與演算法，不容許有遺失值出現

遺失值處理的精髓就在於不能影響整體的估計，讓程式執行下去，讓已有的別的屬性的值能發揮作用。填補數與已有的數的分佈、特徵應符合，不能因為處理的變更導致估計值變化。

## 遺失值特性的分類

遺失值根據遺失的特性，大致上可以分成以下三種情形：

| 遺失類型 | 定義 | 例子 |
|---|---|---|
| 完全隨機缺失 (Missing Completely at Random) | 缺失資料與該變數的真實值無關，與其他變數的數值也無關。 | 老師走在路上，不小一掉了幾張考卷，所以成績的遺失和成績無關，和學生其他變數也無關。 |
| 條件隨機缺失 (Missing at Random) | 缺失資料與其他變數有關。 | 一群學生的體重，發現有些人的體重缺失，後來發現女生不喜歡寫體重資訊，因此體重的資訊遺失與性別相關。 |
| 非隨機缺失 (Missing not an Random) | 缺失資料依賴於該變數本身。 | 有分收入的問卷調查，通常薪資高的人，不喜歡填有關收入資訊，所以收入資料缺失和收入高低有關。 |

## 處理遺失值的方法 - 刪除

處理遺失值的方法，可以分為刪除和補值兩類。

**刪除**

- 刪除有缺失資料的樣本
- 刪除有過多缺失資料的變數 (通常超過 60% 就會刪除)
- 缺點：
  # 會導致資訊丟失。
  # 不是完全隨機缺失，若採用簡單刪除法就會使得估計係數出現偏誤。

## 處理遺失值的方法 - 補值

**補值**

- 給定一個固定值去填補遺失值
- 由後往前補值 (時間性相關適用)
- 由前往後補值 (時間性相關適用)
- 用現有的資料取平均值、中位數、眾數等進行補值
- 用預測方法補值，迴歸或機器學習

## 回到今天的課程範例

運用 python ，偵測與判斷是否有重複或遺失值，以實際資料，如何進行資料遺失值處理。

## Python 語法

- 觀察-是否有重覆

```
#顯示有重覆的資訊:
df_train.duplicated()
```

- 用平均值補值

```
df_train['Age']=df_train['Age'].fillna(df_train['Age'].mean())
```

- 由前往後補植：

```
df_train['Age']=df_train['Age'].fillna(method='bfill')
```

- 由後往前補：

```
df_train['Age']=df_train['Age'].fillna(method='pad')
```

## 知識點回顧

- 以 kaggle 鐵達尼資料集為例，理解資料科學數據議題
- 遺失值有三大種類，包含完全隨機缺失

**條件隨機缺失**與非隨機缺失

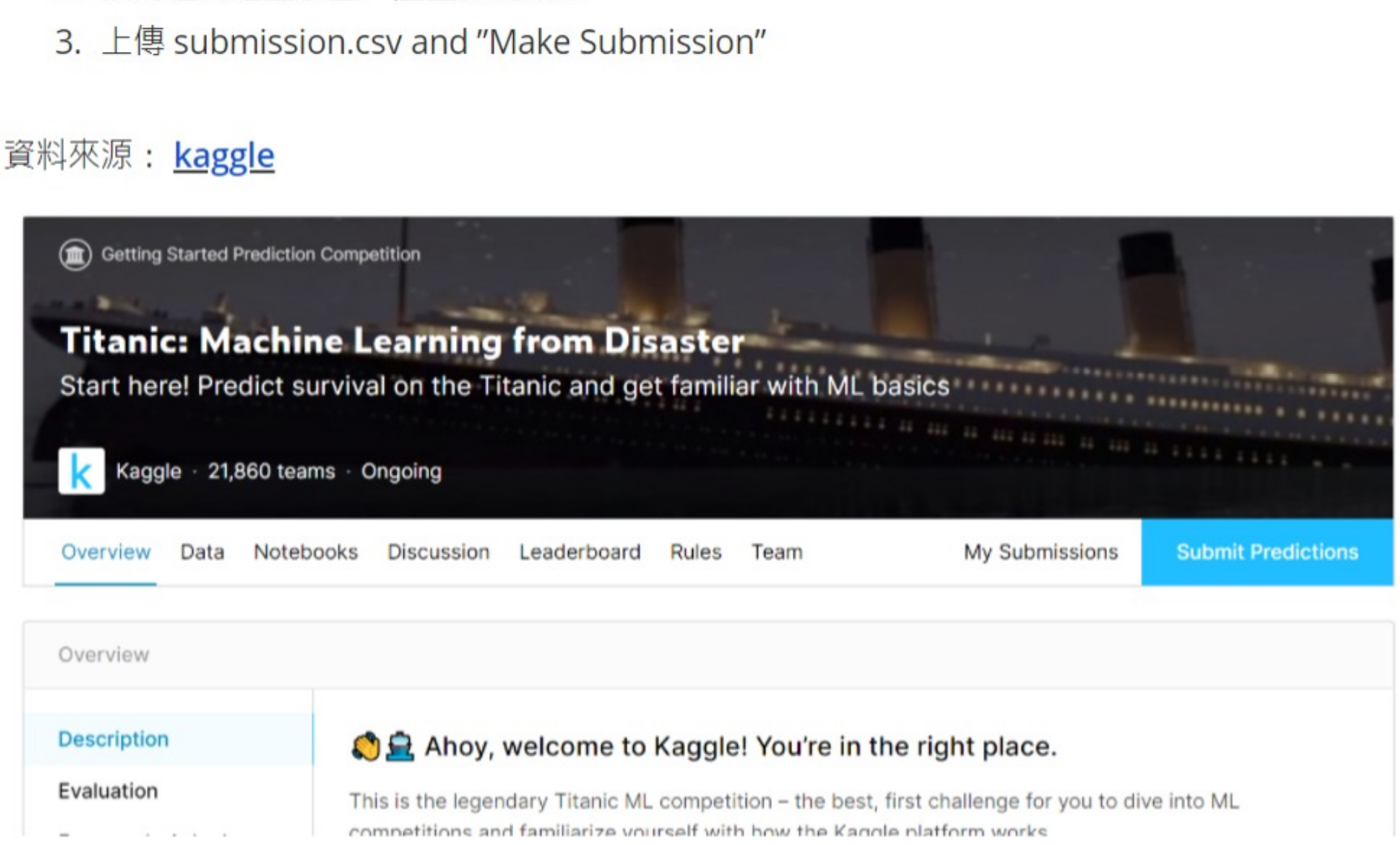
- 運用 python ，處理資料中重覆和遺失值處理，可透過刪除和補值的技巧處理遺失值。

## 延伸閱讀

Kaggle 參與比賽的流程

1. 下載資料(資料與 gender_submission.csv)
2. 撰寫程式建立模型，產生預測結果
3. 上傳 submission.csv and "Make Submission"

資料來源： kaggle



Titanic: Machine Learning from Disaster

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

Kaggle · 21,860 teams · Ongoing

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.

下一步：閱讀範例與完成作業