資料探勘 期中報告第二組

眼睛識別

112029004周珮珊
112029013呂盈萱
112029021林姵安
112029022黃少虹
112029052黃筑暄

指導教授:鮑永誠

November I 2025

目錄

問題描述

題目：年齡對哪個眼睛相關疾病最具影響

由於尹書田醫療財團法人書田泌尿科眼科診所-中年人常見的眼疾及預防保養之道
指出，隨著年齡的增加，患上與年齡相關的眼疾的風險也隨之增加，這些病症包括白內障、青光眼、糖尿病視網膜病變、黃斑病變等。因此，我們想探討年齡對哪個眼睛相關疾病最具影響，而每個人都應該持續關注自己的眼健康，採取主動而非被動的態度。

資料描述

本資料集主要包含眼病智能識別（ODIR）資料，收錄5000名患者的年齡、左右眼彩色眼底照片與醫師診斷關鍵詞。

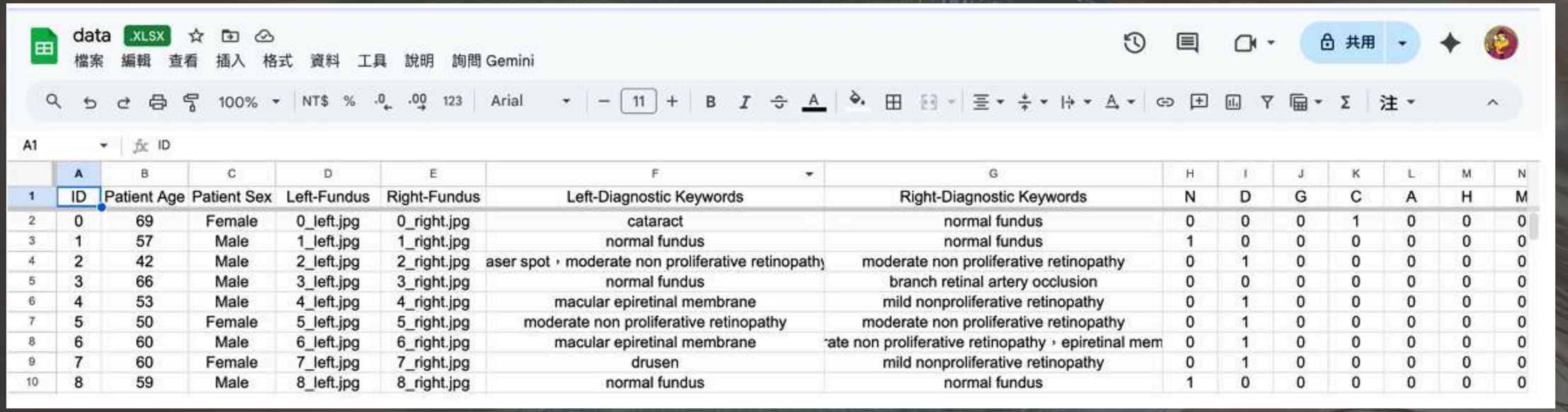
資料代表上工醫療科技有限公司自中國多家醫療機構收集的真實臨床影像，涵蓋多種相機設備（如佳能、蔡司、科和），圖像解析度不一。
標註由專業人員經品質控管後完成，患者依診斷結果分為八類：

正常（N）、糖尿病（D）、青光眼（G）、白內障（C）、年齡相關性黃斑變性（A）、高血壓（H）、病理性近視（M）以及其他疾病／異常（O）。

資料描述

112029022黃少虹

- 6393 Rows
- 19 columns



data .XLSX

檔案 編輯 查看 插入 格式 資料 工具 說明 詢問 Gemini

共用

A1 | ID

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	Patient Age	Patient Sex	Left-Fundus	Right-Fundus	Left-Diagnostic Keywords	Right-Diagnostic Keywords	N	D	G	C	A	H	M
2	0	69	Female	0_left.jpg	0_right.jpg	cataract	normal fundus	0	0	0	1	0	0	0
3	1	57	Male	1_left.jpg	1_right.jpg	normal fundus	normal fundus	1	0	0	0	0	0	0
4	2	42	Male	2_left.jpg	2_right.jpg	aser spot，moderate non proliferative retinopathy	moderate non proliferative retinopathy	0	1	0	0	0	0	0
5	3	66	Male	3_left.jpg	3_right.jpg	normal fundus	branch retinal artery occlusion	0	0	0	0	0	0	0
6	4	53	Male	4_left.jpg	4_right.jpg	macular epiretinal membrane	mild nonproliferative retinopathy	0	1	0	0	0	0	0
7	5	50	Female	5_left.jpg	5_right.jpg	moderate non proliferative retinopathy	moderate non proliferative retinopathy	0	1	0	0	0	0	0
8	6	60	Male	6_left.jpg	6_right.jpg	macular epiretinal membrane	·ate non proliferative retinopathy，epiretinal mem	0	1	0	0	0	0	0
9	7	60	Female	7_left.jpg	7_right.jpg	drusen	mild nonproliferative retinopathy	0	1	0	0	0	0	0
10	8	59	Male	8_left.jpg	8_right.jpg	normal fundus	normal fundus	1	0	0	0	0	0	0

資料描述

112029022黃少虹

連續性

欄位	資料型態	定義
Age	等比（數值）	患者年齡

類別型

欄位	資料型態	定義
ID	順序（數值）	患者編號
Patient Sex	名目（文字）	患者性別
Left-Fundus	（圖片）	左眼彩色底照
Right-Fundus	（圖片）	右眼彩色底照
Left-Diagnostic Keywords	名目（文字）	左眼診斷關鍵字
Right-Diagnostic Keywords	名目（文字）	右眼診斷關鍵字

資料描述

112029022黃少虹

- 0:無得該病、1：有得該病

類別型

欄位	資料型態	定義
N	名目（數值）	正常
D	名目（數值）	糖尿病
G	名目（數值）	青光眼
C	名目（數值）	白內障
A	名目（數值）	年齡相關性黃斑變性
H	名目（數值）	高血壓
M	名目（數值）	病理性近視
O	名目（數值）	其他疾病/異常

資料描述

112029022黃少虹

data1 .XLSX

檔案 編輯 查看 插入 格式 資料 工具 說明 ...

100% NT$ % .0 .00 123 Arial

A1 | ID

	A	B	C	D	E	F	G	H	I	J
1	ID	Patient Age	N	D	G	C	A	H	M	O
2	0	69	0	0	0	1	0	0	0	0
3	1	57	1	0	0	0	0	0	0	0
4	2	42	0	1	0	0	0	0	0	1
5	3	66	0	0	0	0	0	0	0	1
6	4	53	0	1	0	0	0	0	0	1
7	5	50	0	1	0	0	0	0	0	0
8	6	60	0	1	0	0	0	0	0	1
9	7	60	0	1	0	0	0	0	0	1
10	8	59	1	0	0	0	0	0	0	0

- 3500 Rows
- 10 columns

- 只保留ID、年齡、疾病代碼的欄位 ，其他手動刪除

一 資料前處理

- 導入檔案

Step1. 導入所需函式庫

Step2. 設定資料檔案的路徑變數 data_path = "data.xlsx"、使用 if os.path.exists(data_path): 來確認資料檔案是否存在於預期的位置

```
# ========================================
# Step 1. Import Packages
# ========================================
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
import seaborn as sns
import os

# ========================================
# Step 2. Load Data (Excel)
# ========================================
data_path = "data.xlsx"
if os.path.exists(data_path):
    df = pd.read_excel(data_path)
    print(f" File loaded: Number of records = {len(df)}")
else:
    print(" File not found, please check data.xlsx exists.")
    df = None

if df is not None:
```

Step3. 缺值處理 (Missing Value Processing)

1. 若「年齡」有缺失值 → 以該欄位的算術平均值填補
2. 若疾病欄位 (N, D, G, C, A, H, M, O) 有缺值 → 補 0（視為「沒有疾病」）

年齡算術平均數：58.85

```
    # ================================================
    # Step 3. Missing Value Processing
    # ================================================
    df["Patient Age"] = df["Patient Age"].fillna(df["Patient
Age"].mean())
    disease_cols = ["N", "D", "G", "C", "A", "H", "M", "O"]
    for col in disease_cols:
        df[col] = df[col].fillna(0)
```

一 資料前處理

Step4. 篩選特定樣本 (Filtering)

只保留「N=0」的樣本資料

```
# ============================================
# Step 4. Filter N=0
# ============================================
df = df[df["N"] == 0]
print(f" After filtering N=0, record count: {len(df)}")
```

資料筆數：3500➡2360

```
After filtering N=0, record count: 2360
```

一 資料前處理

Step5. 移除異常值 (Outlier Removal)

- 計算年齡的平均值與標準差
- 將「超過平均 ±3個標準差」的年齡視為異常，並排除

資料筆數：2360➡2343

標準差：10.96
上界：91.74
下界：25.97

```
# ============================================
# Step 5. Outlier Processing (Age)
# ============================================
mean_age = df["Patient Age"].mean()
std_age = df["Patient Age"].std()
lower = mean_age - 3 * std_age
upper = mean_age + 3 * std_age
df = df[(df["Patient Age"] >= lower) & (df["Patient Age"] <= upper)]
print(f" After removing outliers, record count: {len(df)}")
```

```
After removing outliers, record count: 2343
```

一 資料前處理

一 資料前處理

Step6. 疾病欄位編碼

將疾病欄位的值統一轉成 0 或 1

Step7. 年齡標準化

使用 Z-score將年齡轉換成[0,1]

```
# ============================================
# Step 6. Disease column encoding
# ============================================
for col in disease_cols:
    df[col] = df[col].apply(lambda x: 1 if x == 1 else 0)


# ============================================
# Step 7. Age normalization (Z-score)
# ============================================
scaler = StandardScaler()
df["Patient Age_z"] = scaler.fit_transform(df[["Patient Age"]])
```

一 資料前處理

Step8. 依年齡劃分兩組
（Adult成人 / Elderly老年)

- 依年齡分組

"Adult (15-64)" 有 1603 筆
"Elderly (65+)" 有 740 筆

```
# ==========================================
# Step 8. Age Grouping
# ==========================================
def age_group(age):
    if 0 <= age <= 14:
        return 'Juvenile (0-14)'
    elif 15 <= age <= 64:
        return 'Adult (15-64)'
    else:
        return 'Elderly (65+)'
df['Age_Group'] = df['Patient Age'].apply(age_group)
print("Age group statistics:")
print(df['Age_Group'].value_counts())
```

```
Age_Group
Adult (15-64)    1603
Elderly (65+)     740
Name: count, dtype: int64
```

資料描述

112029004周珮珊

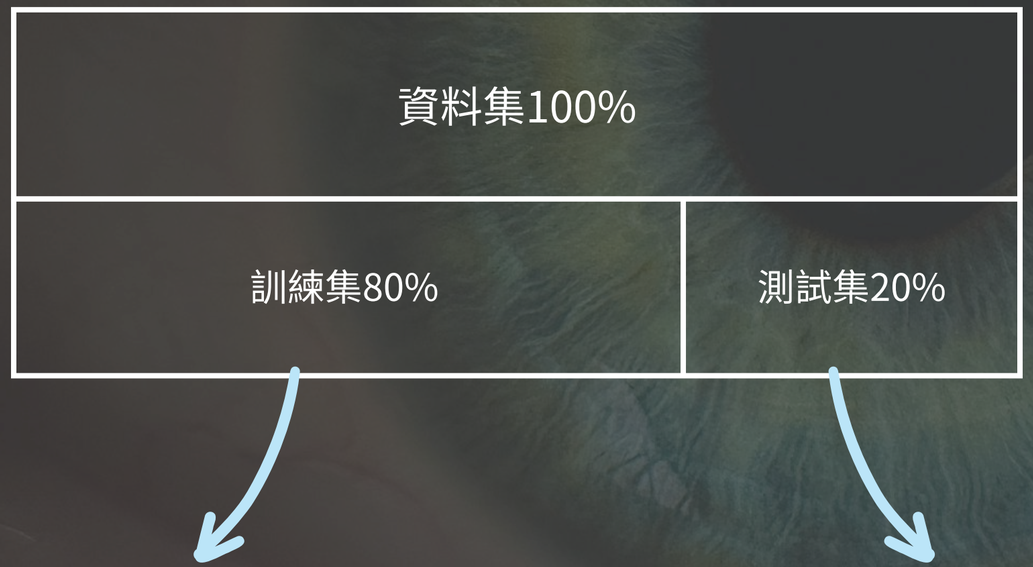
- 2343 Rows
- 11 columns

D1 | N

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Patient Age	atient Age	Age_Grou	N	D	G	C	A	H	M	O		
2	69	0.926	Elderly (6	0	0	0	1	0	0	0	0		
3	42	-1.5377	Adult (15-	0	1	0	0	0	0	0	1		
4	66	0.6523	Elderly (6	0	0	0	0	0	0	0	1		
5	53	-0.534	Adult (15-	0	1	0	0	0	0	0	1		
6	50	-0.8077	Adult (15-	0	1	0	0	0	0	0	0		
7	60	0.1048	Adult (15-	0	1	0	0	0	0	0	1		
8	60	0.1048	Adult (15-	0	1	0	0	0	0	0	1		
9	54	-0.4427	Adult (15-	0	0	0	0	0	0	0	1		
10	70	1.0173	Elderly (6	0	0	0	0	0	0	0	1		

實驗設計

- 資料切分



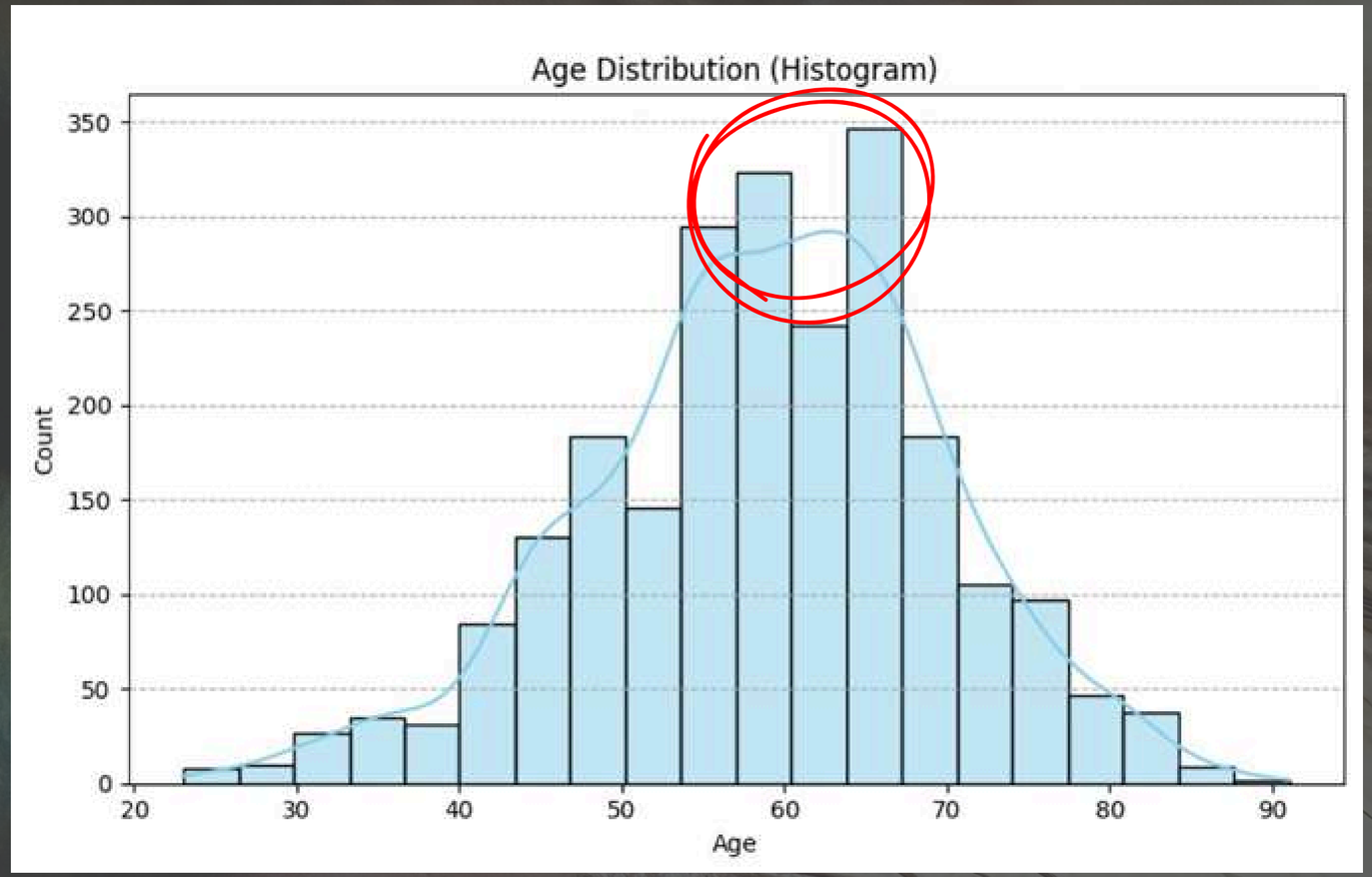
- 交叉驗證

五折交叉驗證

每次用 4 份當訓練、1 份當驗證，輪 5 次，把 5 次的評分平均作為模型效能估計。

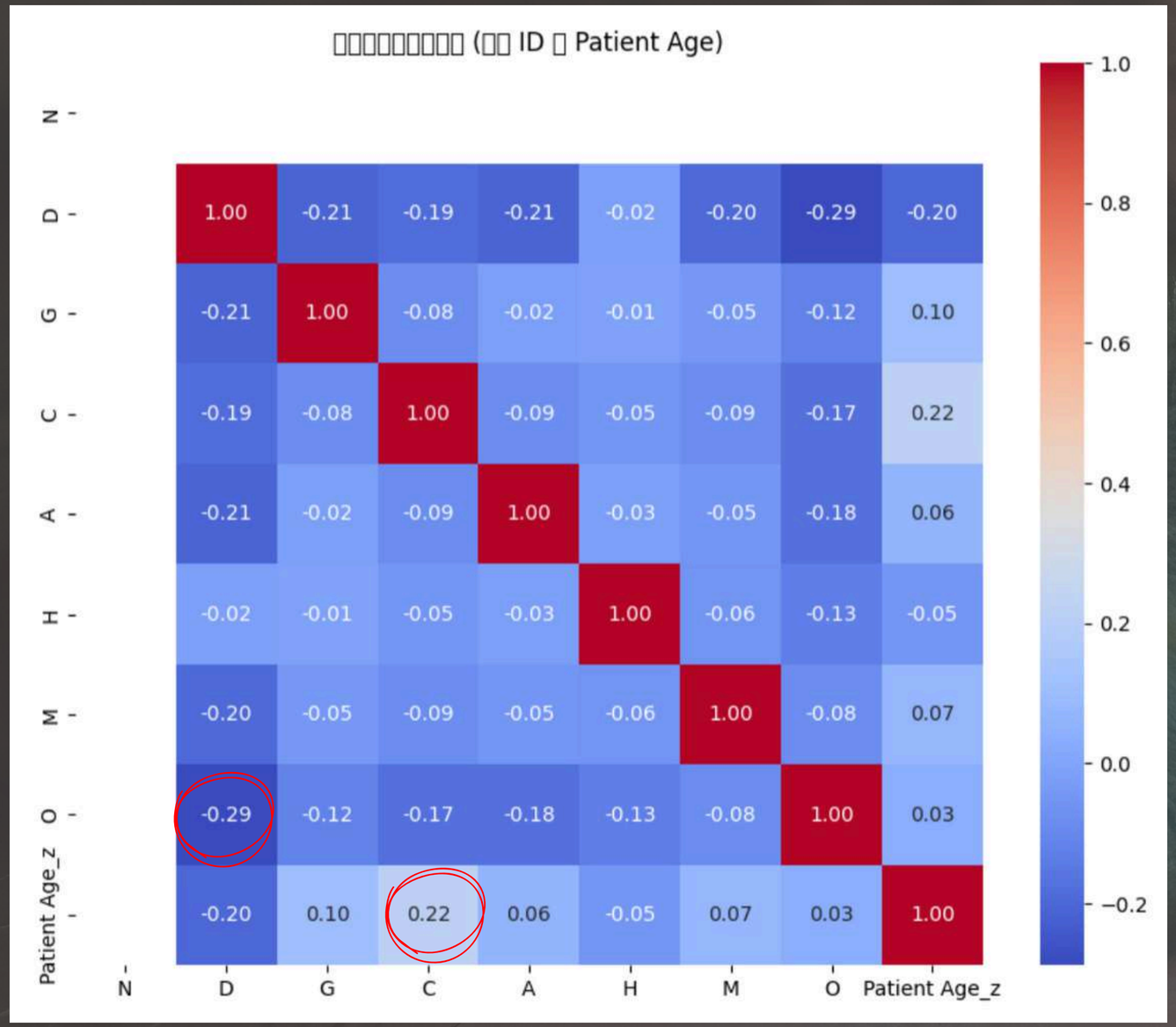
視覺化分析- 直方圖

- 樣本量高峰在 55-70 歲區間，｜離散度左右大致對稱｜左尾微長。這種年齡結構常見於中高齡族群調查資料。



112029004周珮珊

視覺化分析-熱力圖



越紅 → 正相關（+1 代表完全正相關）
越藍 → 負相關（-1 代表完全負相關）
接近 0 → 幾乎沒有線性關係

年齡越大，越有可能罹患白內障
但0.22<0.3（低度相關）

112029004周珮珊

研究議題&預期結果

Patient Age 與疾病變數 (D, G, C, A, H, M, O) 的相關性都很低（大多在±0.2以內） 表示「年齡」與「是否患該疾病」之間沒有明顯的線性關係。

疾病之間的相關性也普遍偏低（多數在 -0.2 ~ 0.2 之間） 顯示這些疾病之間彼此的發生沒有太強的關聯性。 例如：D 與 O 之間的相關係數為 -0.29 → 稍微負相關（可能一種疾病出現時另一種較少見）。

參考文獻

題目的參考文獻：

- https://happy50plus.org/2025/02/eye-problems-related-to-age/
- https://reurl.cc/rK1m1O

年齡如何分組的參考文獻：

- https://www.ehanlin.com.tw/app/keyword/國中/地理/年齡組成.html

分工表

周珮珊	做簡報、報告
呂盈萱	視覺化分析、簡報排版
林姵安	簡報製作、文獻探討
黃少虹	資料前處理、報告
黃筑暄	資料前處理、實驗設計

提問Q&A

Q1.如果要針對年齡分組應該先做此步驟再開始前處理嗎？
A：先前處理，但題目要修改，例如共病或風險傳遞類型的關聯分析，因為題目變相不多，比較適合做疾病之間是否有相關性的題目。

Q2.針對「年齡對哪個眼睛相關疾病最具影響」還可以新增哪些視覺化分析?
A：熱力圖很多變項，建議可使用CNN做圖像分析。

THANK YOU FOR LISTENING