從數據看肺癌風險

25國家之人口特徵與環境因子分析

報告者:李芳珊

專題說明

- 本研究專題採用 Kaggle 所提供之「Lung Cancer Risk in 25 Countries」資料集，包含 220,632 筆樣本資料，涵蓋 23 個欄位，記錄來自 25 個國家的個體資料。數據內容涉及人口背景、吸菸習慣、環境暴露、家族病史以及肺癌診斷情形等資訊。根據國家發展程度劃分，其中來自已開發國家的資料占 23.97%，開發中國家則占 76.03%；若依地理區域分類，則包括亞洲 43.96%、歐洲 23.98%、非洲 20.10%、北美洲 7.98%、以及南美洲 3.99%。

  - 資料來源: https://www.kaggle.com

分析目的

1.透過分析 25 國肺癌資料，找出影響肺癌風險的主要因素。

2.比較不同族群（性別、年齡、吸菸、家族史）的風險差異。

3.利用 Power BI 、python進行視覺化與統計分析，挖掘潛在高風險族群。

4.提供數據支持，作為肺癌預防與健康政策的參考依據。

檔案23欄位說明(1)

序號	欄位名稱	欄位說明
1	ID	每筆紀錄的唯一識別碼。
2	Country	資料收集地或該個體所居住的國家。
3	Population Size	研究中所涵蓋國家或地區的總人口數。
4	Age	個體的年齡（單位：歲）。
5	Gender	個體的性別（例如：男性、女性）。
6	Smoker	是否為吸菸者（是/否）。
7	Years of Smoking	該個體吸菸的總年數。
8	Cigarettes per Day	每天平均吸菸的根數。
9	Passive Smoker	是否有暴露於二手菸（是/否）。
10	Family History	是否有肺癌家族病史（是/否）。
11	Lung Cancer Diagnosis	是否已被診斷為肺癌（是/否）。

檔案23欄位說明(2)

序號	欄位名稱	欄位說明
12	Cancer Stage	診斷後的癌症期別（第1至第4期，未確診者為 NaN）。
13	Survival Years	診斷後的追蹤存活年數（未確診者為 0，確診者範圍為 1–10 年）。
14	Adenocarcinoma Type	是否為腺癌類型（是/否）。
15	Air_Pollution_Exposure	空氣污染暴露程度，分類為：中、低、高。
16	Occupational Exposure	是否有職業性暴露（是/否）。
17	Indoor Pollution	是否有室內污染暴露（是/否）。
18	Healthcare_Access	醫療可近性，分為：良好/差。
19	Early_Detection	是否有早期偵測（是/否）。
20	Developed or Developing	國家層級判定為已開發或開發中國家（是/否）。
21	Annual_Lung_Cancer_Deaths	每年肺癌死亡人數（以國家為單位的整數值）。
22	Lung_Cancer_Prevalence_Rate	肺癌盛行率（以國家為單位的浮點數）。
23	Mortality Rate	死亡率（未確診者為 NaN，為浮點數）。

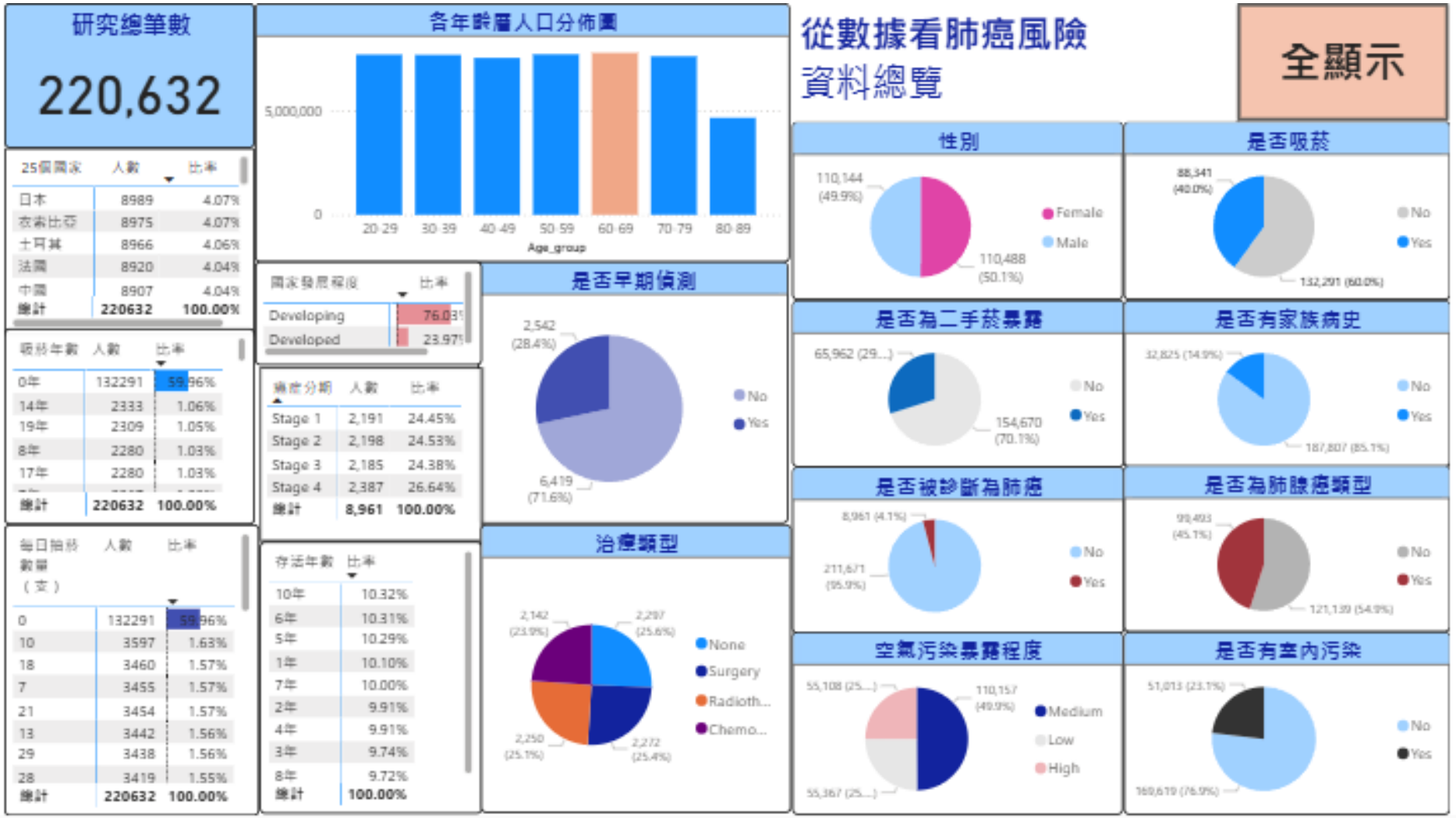
類別變項統計摘要

欄位名稱	分類	最大占比
性別	Female / Male	Female (50.1%)
吸菸	No / Yes	No (60.0%)
二手菸暴露	No / Yes	No (70.1%)
家族病史	No / Yes	No (85.1%)
診斷為肺癌	No / Yes	No (95.9%)
國家發展程度	Developing / Developed	Developing (76.0%)
醫療可近性	Poor / Good	Poor (80.0%)
是否早期偵測	No / Yes	No (72.0%)

Microsoft Power BI

性別	吸菸人數	非吸菸人數	吸菸率	非吸菸率
Male	60,687	49,457	55.10%	44.90%
Female	27,654	82,834	25.03%	74.97%
總計	88,341	132,291	40.04%	59.96%

Microsoft Power BI

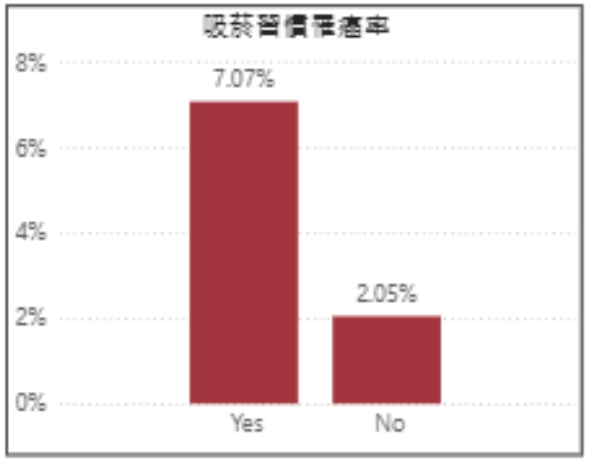


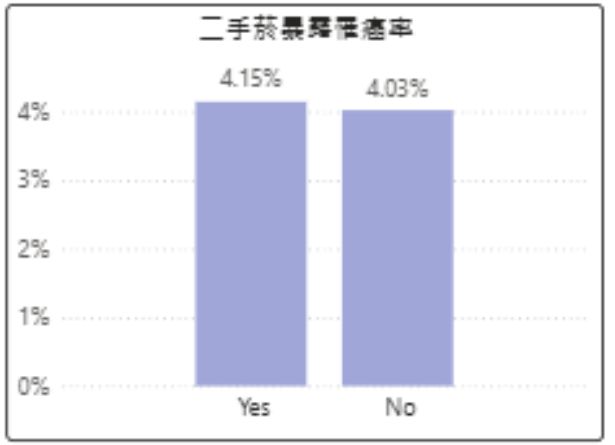
Lung Cancer Risk in 25 Countries_資料總覽(2)

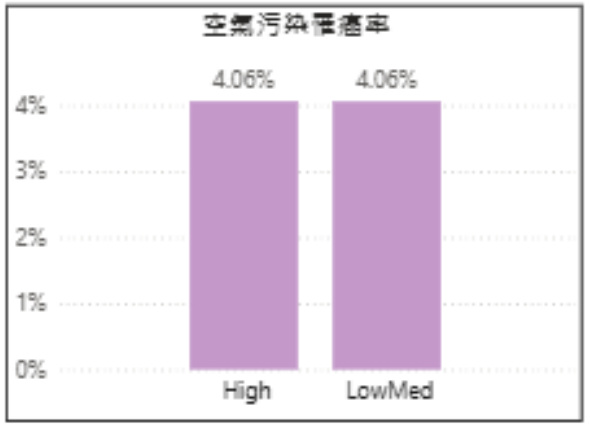
即時資料 資料已於 25/7/27 下午10:07更新

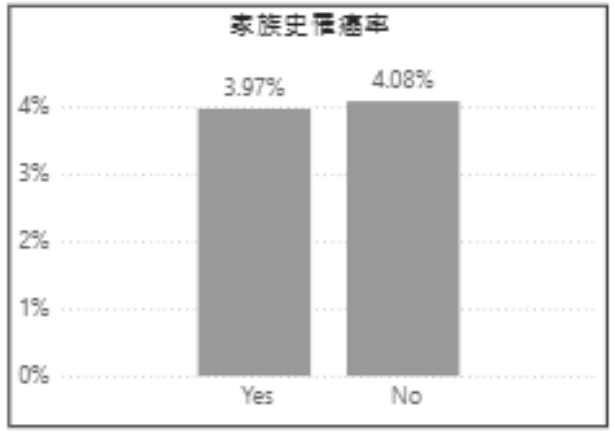
Microsoft Power BI

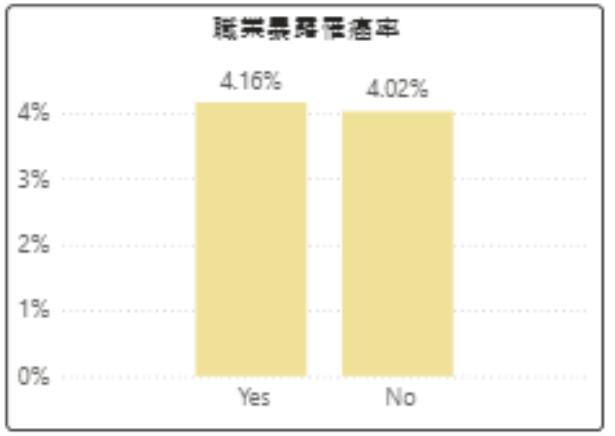
風險因子分析_與肺癌診斷之相關性











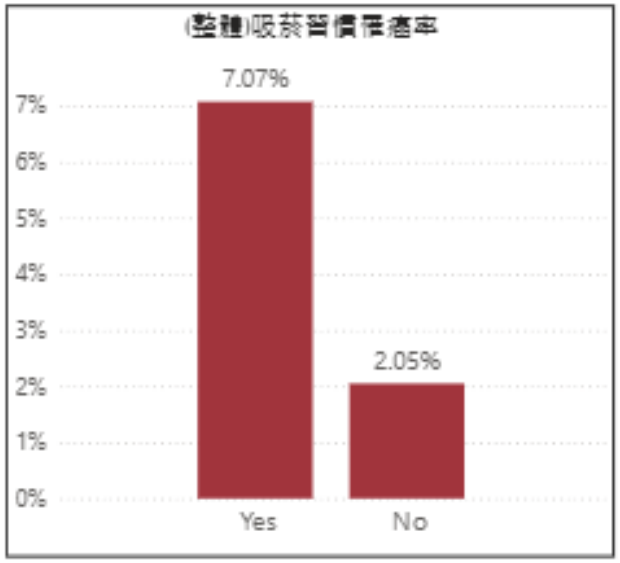
篩選

Lung Cancer Risk in 25 Countries_風險因子分析_與肺癌診斷之相關性

即時資料

資料已於 25/7/27 下午10:07更新

Microsoft Power BI



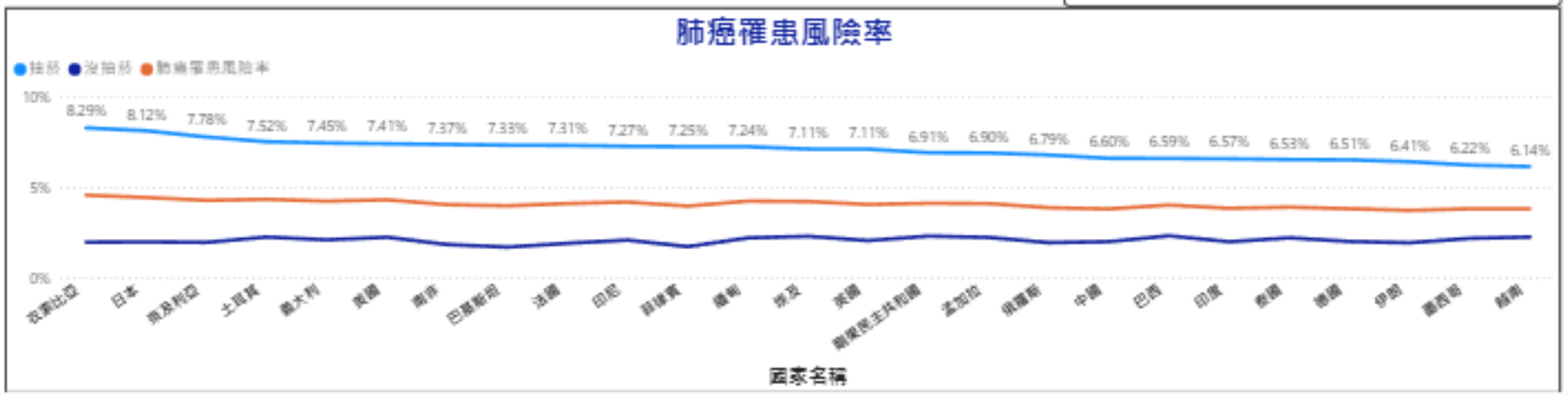
抽菸肺癌罹癌率

7.07%

沒抽菸肺癌罹癌率

2.05%

國家名稱	抽菸	沒抽菸
衣索比亞	8.29%	1.95%
日本	8.12%	1.96%
奈及利亞	7.78%	1.94%
土耳其	7.52%	2.23%
義大利	7.45%	2.08%
美國	7.41%	2.22%
南非	7.37%	1.82%
巴基斯坦	7.33%	1.68%
法國	7.31%	1.89%
印尼	7.27%	2.07%
菲律賓	7.25%	1.70%
總計	7.07%	2.05%

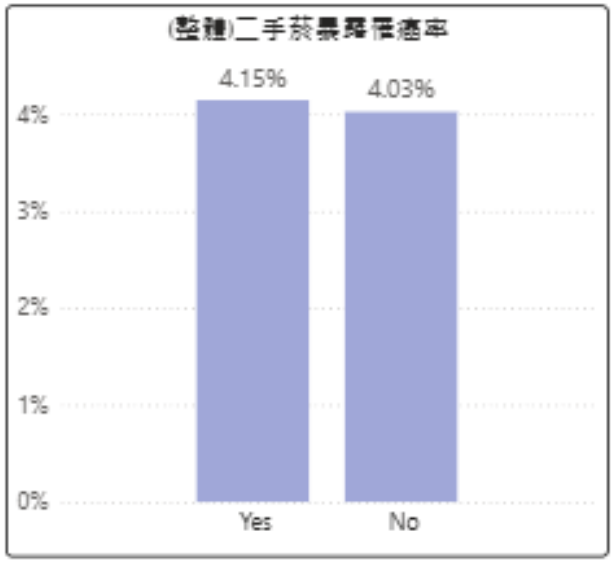


Lung Cancer Risk in 25 Countries, 吸菸習慣肺癌罹癌率

即時資料

資料已於 25/7/27 下午10:07更新

Microsoft Power BI

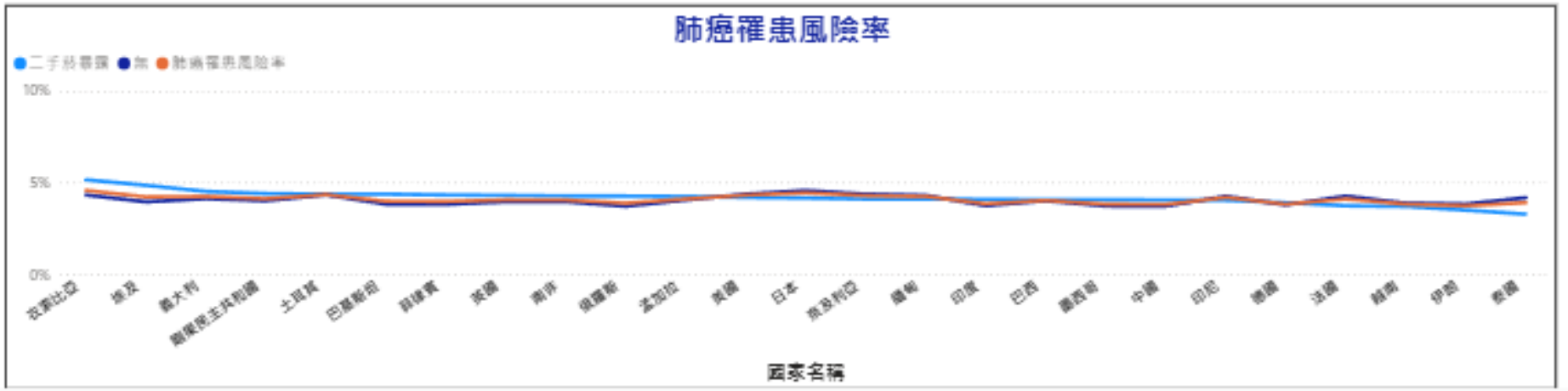


二手菸暴露罹癌率

無二手菸暴露罹癌率

4.03%

國家名稱	二手菸暴露	無
衣索比亞	5.14%	4.31%
埃及	4.84%	3.93%
義大利	4.50%	4.11%
剛果民主共和國	4.39%	3.99%
土耳其	4.35%	4.32%
巴基斯坦	4.35%	3.81%
菲律賓	4.31%	3.80%
英國	4.29%	3.94%
南非	4.26%	3.94%
俄羅斯	4.25%	3.69%
委加拉	4.23%	4.03%
總計	4.15%	4.03%

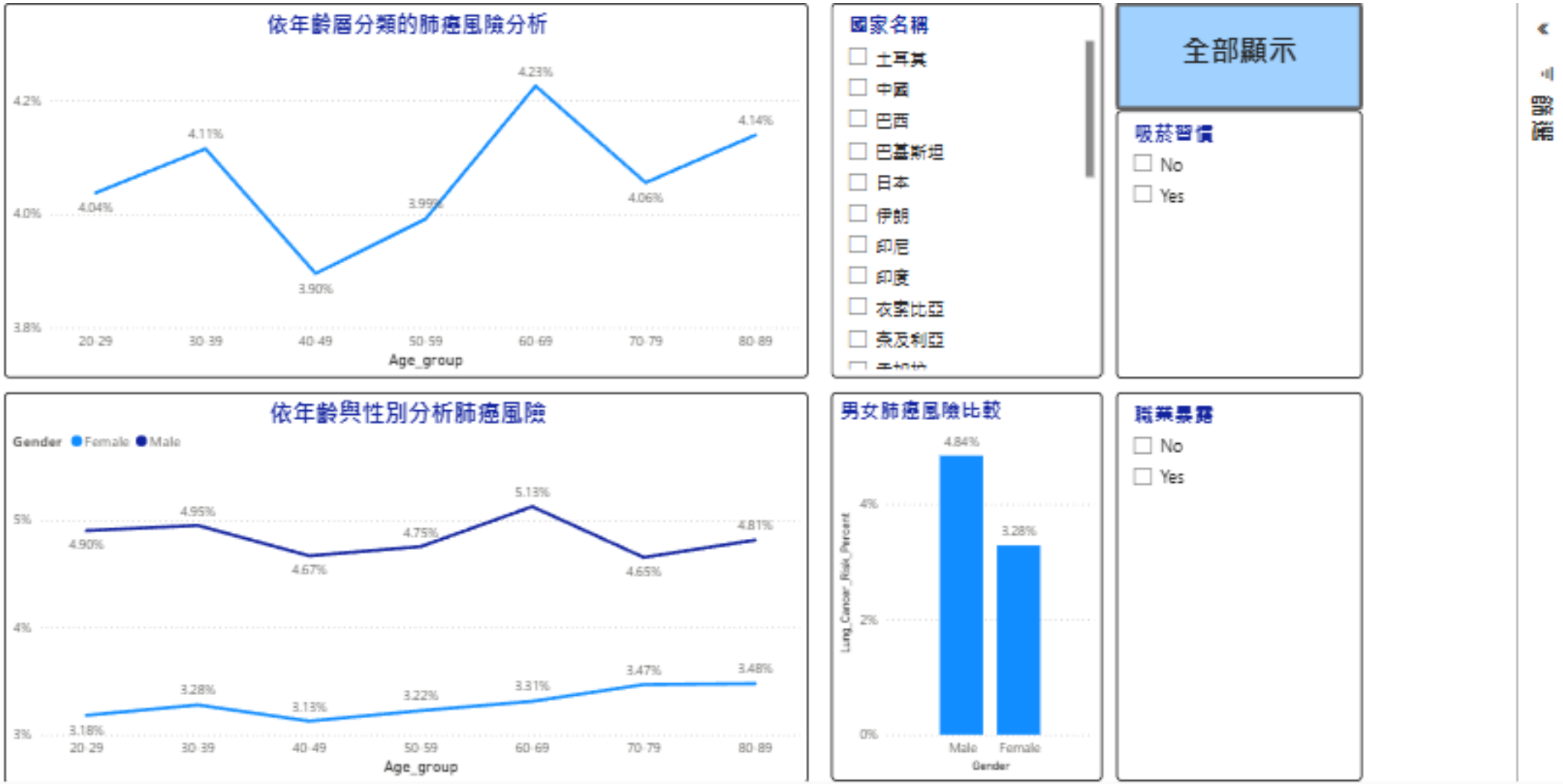


Lung Cancer Risk in 25 Countries, 二手菸暴露罹癌率

即時資料

資料已於 25/7/27 下午10:07更新

Microsoft Power BI



Microsoft Power BI

國家發展程度

發展程度	次數	比率
Developed	52,891	23.97%
Developing	167,741	76.03%
總計	**220,632**	**100.00%**

確診肺癌_國家發展程度

發展程度	次數	比率
Developed	2,195	24.50%
Developing	6,766	75.50%
總計	**8,961**	**100.00%**

各期別_存活率

Stage	1年	2年	3年	4年	5年	6年	7年	8年	9年	10年	總計
Stage 1	10.91%	9.77%	9.86%	10.82%	9.04%	10.68%	10.13%	9.63%	10.04%	9.13%	**100.00%**
Stage 2	9.33%	9.60%	9.60%	9.33%	10.96%	10.65%	9.92%	9.65%	9.96%	11.01%	**100.00%**
Stage 3	9.89%	10.11%	9.02%	9.47%	11.17%	9.98%	10.02%	9.84%	9.52%	10.98%	**100.00%**
Stage 4	10.26%	10.14%	10.43%	10.01%	10.01%	9.97%	9.93%	9.76%	9.30%	10.18%	**100.00%**
總計	**10.10%**	**9.91%**	**9.74%**	**9.91%**	**10.29%**	**10.31%**	**10.00%**	**9.72%**	**9.70%**	**10.32%**	**100.00%**

吸菸年數(已確診肺癌)

Years	No	Yes
0年	30.26%	
1年		1.58%
2年		1.82%
3年		1.83%
4年		1.60%
5年		1.72%
6年		1.76%
7年		1.64%
8年		1.57%
9年		2.00%
10年		1.75%
11年		2.04%
12年		1.80%
13年		1.81%
14年		1.90%
15年		1.99%
16年		1.71%
17年		1.91%
18年		1.50%
19年		1.60%
總計	**30.26%**	**69.74%**

描述統計_確診是肺癌(1)

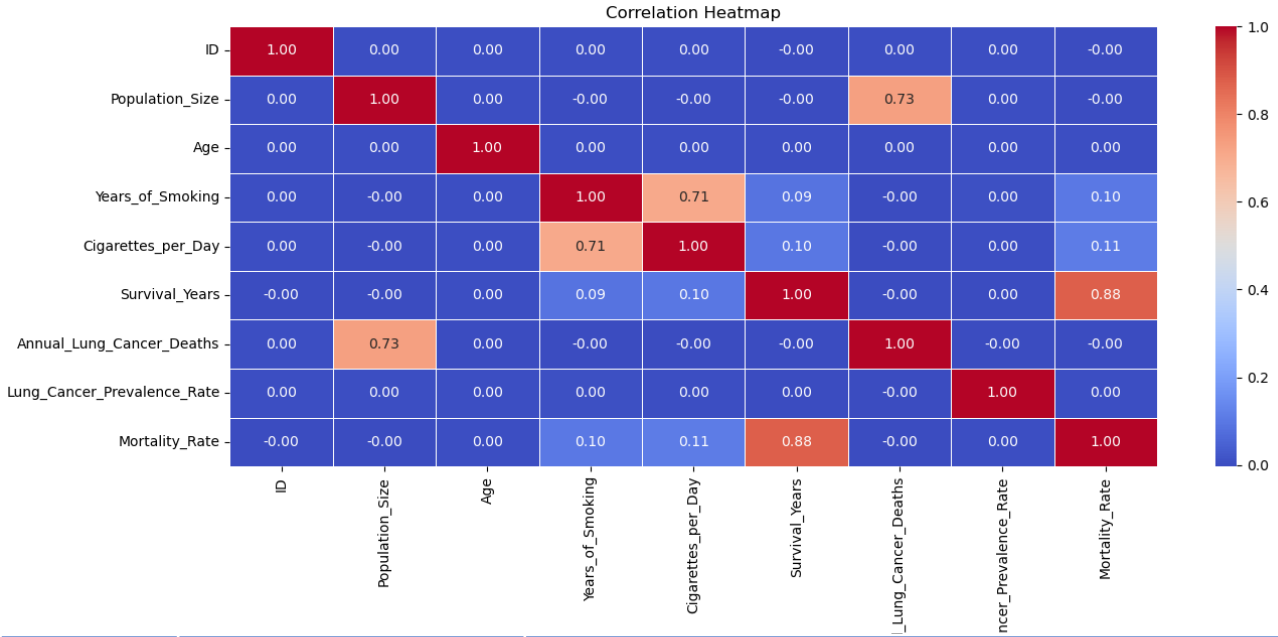
肺癌確診者 8,961	人口規模	年齡	吸菸年數	每日抽菸數量（支）	存活年數	年肺癌死亡人數	肺癌盛行率%	死亡率%
mean	224.20	52.66	14.24	12.21	5.50	62,356	1.51	75.09
std	339.62	19.18	13.43	10.17	2.87	126,716	0.58	8.62
min	54	20	0	0	1	10,044	0.5	60
25%	83	36	0	0	3	23,000	1.01	67.57
50%	113	53	12	12	6	30,000	1.5	75.08
75%	206	69	26	21	8	45,000	2.01	82.61
max	1400	85	40	30	10	690,000	2.5	90

描述統計_確診是肺癌(2)

年齡_按性別描述統計			
肺癌確診者	年齡	男/年齡	女/年齡
count	8,961	5,332	3,629
mean	52.66	52.33	53.16
std	19.18	19.09	19.29
min	20	20	20
25%	36	35	36
50%	53	52	54
75%	69	69	70
max	85	85	85

存活年數_按癌症期別_描述統計					
肺癌確診者存活年數	存活年數	Stage 1	Stage 2	Stage 3	Stage 4
count	8,961	2,191	2,198	2,185	2,387
mean	5.50	5.42	5.60	5.55	5.45
std	2.87	2.87	2.86	2.88	2.88
min	1	1	1	1	1
25%	3	3	3	3	3
50%	6	5	6	6	5
75%	8	8	8	8	8
max	10	10	10	10	10

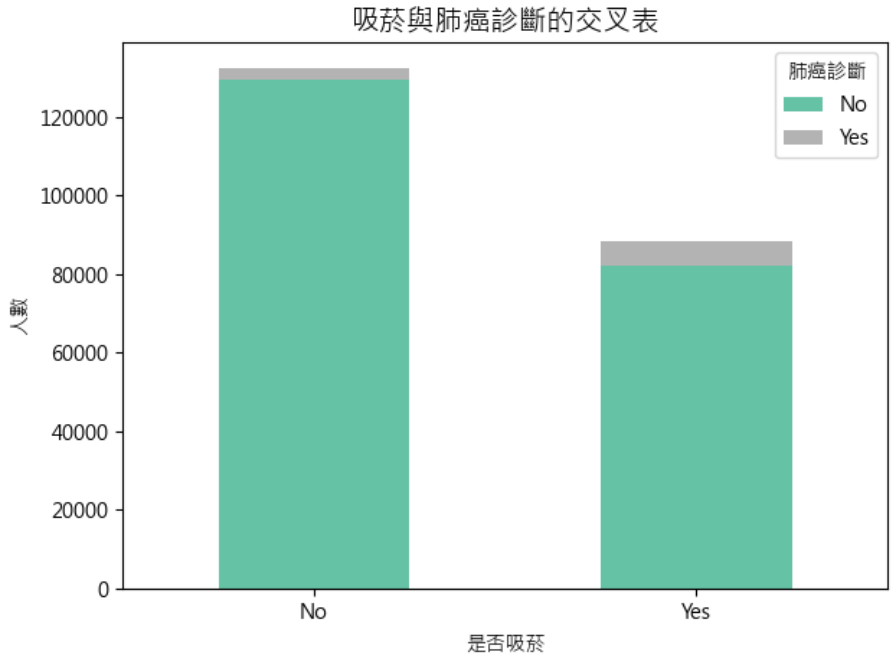
熱力圖

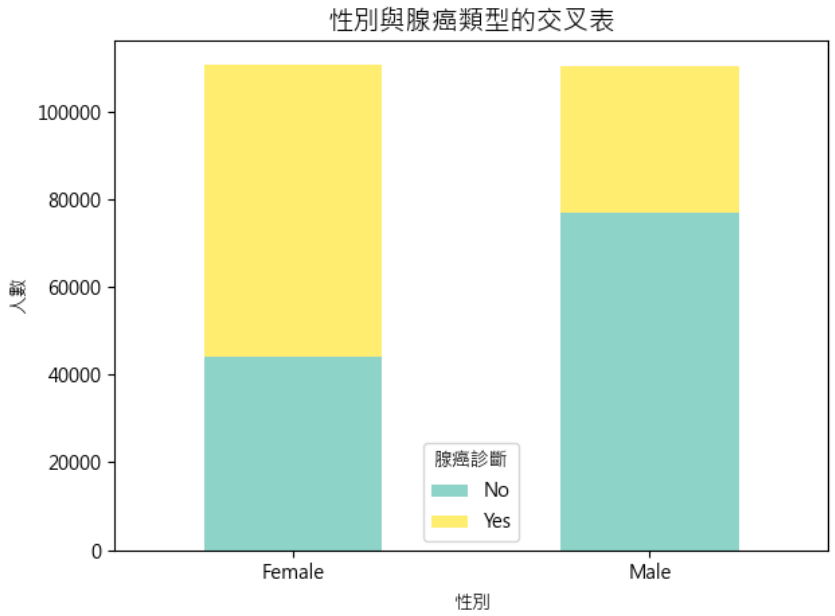


變數組合	相關係數	相關性類型	解釋說明
吸菸年數與每天吸菸量	0.71	正相關	吸菸年數愈久的人，平均每天吸菸數量也愈多。
存活年數與死亡率	0.88	正相關	國家若有高齡人口比例，死亡率可能偏高
人口數與每年肺癌死亡人數	0.73	正相關	人口多的國家，肺癌死亡人數也多。
年齡與其他變項	0.00	無明顯線性相關	「年齡」變化與其他變項（如吸菸、癌症死亡率）無明顯線性關係

肺癌診斷與吸菸 / 性別關聯分析

- **吸菸與肺癌診斷**：
  Cramér’ s V = **0.125** → 表示有**輕微的關聯**，但並不強烈，可能還需搭配其他風險因子分析。
- **性別與腺癌類型**：
  Cramér’ s V = **0.301** → 表示有**中度以上的關聯性**，說明性別可能在特定肺癌型態（如腺癌）上有一定的關聯分佈。





卡方檢定

變項	卡方統計量	自由度	p值	顯著與否(p<0.05)
Smoker	3429.47	1	0.0000	顯著
Passive_Smoker	1.71	1	0.1915	不顯著
Family_History	0.87	1	0.3523	不顯著
Air_Pollution_Exposure	0.42	2	0.8115	不顯著
Occupational_Exposure	2.26	1	0.1331	不顯著
Indoor_Pollution	1.07	1	0.3014	不顯著

吸菸（Smoker）」是唯一對「肺癌診斷（Lung_Cancer_Diagnosis）」有**顯著關係**的因素，其餘如二手菸暴露、家族病史、空氣污染、職業暴露、室內污染皆未達統計顯著。

雙樣本獨立 t 檢定_依肺癌診斷分組

變項名稱	t 值	p 值	統計意義
年齡	0.738	0.4608	不具統計顯著性（$p \geq 0.05$）
每日抽菸數量（支）	49.517	0.0000	具統計顯著性（$p < 0.05$）
吸菸年數	43.777	0.0000	具統計顯著性（$p < 0.05$）

羅吉斯迴歸(1)

項目	說明
Dep. Variable: Cancer	依變數（目標）：是否罹患肺癌（1=是, 0=否）
No. Observations: 220632	使用的資料總筆數為 220,632 筆
Model: Logit	模型類型為邏輯斯迴歸（Logistic Regression）
Method: MLE	估計方法為最大概似估計（Maximum Likelihood Estimation）
Pseudo R-squ.: 0.04498	類似決定係數，模型解釋力約為 4.5%
Log-Likelihood: -35798	對數概似函數值，數值越大表示模型越好
LLR p-value: 0.000	整體模型是顯著的（$p < 0.05$）
converged: True	模型成功收斂，結果可信

羅吉斯迴歸(2)

變數名稱	係數（coef）	標準誤（std err）	z 值	p 值	95%信賴區間	勝算比（OR）	中文解釋與結論
const	-3.9059	0.042	-93.79	0.000	[-3.987, -3.824]	0.0201	截距項（基準情況）
Age	0.0004	0.001	0.617	0.537	[-0.001, 0.001]	1.0003	無顯著影響（p > 0.05）
Smoker	1.2910	0.023	55.12	0.000	[1.245, 1.337]	3.6365	✅顯著，吸菸者罹癌風險為非吸菸者的 3.64 倍
Passive Smoker	0.0264	0.024	1.12	0.263	[-0.020, 0.073]	1.0268	不顯著
Family History	-0.0290	0.031	-0.94	0.347	[-0.089, 0.031]	0.9714	不顯著
Air Pollution Exposure	0.0045	0.015	0.29	0.770	[-0.026, 0.035]	1.0045	不顯著
Occupational Exposure	0.0357	0.024	1.52	0.130	[-0.010, 0.082]	1.0363	接近但未達顯著
Indoor Pollution	-0.0236	0.026	-0.91	0.364	[-0.074, 0.027]	0.9767	不顯著
Healthcare Access	0.0174	0.029	0.60	0.548	[-0.039, 0.074]	1.0175	不顯著
Early Detection	0.0130	0.026	0.50	0.615	[-0.038, 0.064]	1.0131	不顯著

結論

- **吸菸是罹患肺癌的最大風險因子**，戒菸是降低肺癌風險最有效的方法。
- 國民健康署（國健署）提供**低劑量電腦斷層掃描（LDCT）肺癌篩檢**服務給符合特定條件的高風險族群。**每兩年一次**，且費用由國健署補助。

肺癌篩檢宣導

- 胸部低劑量電腦斷層攝影檢查(LDCT)
- 補助對象為肺癌高風險族群；
- 若有吸菸情形，應同意接受戒菸服務。

謝 謝 聆 聽