# Netflix 用戶資料的分析報告

Chao Wen Tsai

2025 年 3 月 22 日

## 一、專案概述

本計畫旨在分析 Netflix 用戶數據，以了解用戶的年齡分佈、不同國家的用戶數量、不同訂閱類型的平均觀看時長、最受歡迎的電影類型以及每月登入次數的變化趨勢。

## 二、資料導入與基本資訊查看

程式碼實作 **library(readr)**

## 載入數據

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
data <- read_csv("C:/Users/chao/Desktop/Data Analysis/Datasets/netflix_users.csv")
```

```
## Rows: 25000 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): Name, Country, Subscription_Type, Favorite_Genre
## dbl  (3): User_ID, Age, Watch_Time_Hours
## date (1): Last_Login
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
cat('數據基本資料: \n')
```

```
## 數據基本資料:
```

```r
str(data)
```

```
## spc_tbl_ [25,000 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ User_ID          : num [1:25000] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Name             : chr [1:25000] "James Martinez" "John Miller" "Emma Davis" "Emma Miller" ...
##  $ Age              : num [1:25000] 18 23 60 44 68 21 57 68 39 55 ...
##  $ Country          : chr [1:25000] "France" "USA" "UK" "USA" ...
##  $ Subscription_Type: chr [1:25000] "Premium" "Premium" "Basic" "Premium" ...
##  $ Watch_Time_Hours : num [1:25000] 80.3 321.8 35.9 261.6 909.3 ...
##  $ Favorite_Genre   : chr [1:25000] "Drama" "Sci-Fi" "Comedy" "Documentary" ...
##  $ Last_Login       : Date[1:25000], format: "2024-05-12" "2025-02-05" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   User_ID = col_double(),
##   ..   Name = col_character(),
##   ..   Age = col_double(),
##   ..   Country = col_character(),
##   ..   Subscription_Type = col_character(),
##   ..   Watch_Time_Hours = col_double(),
##   ..   Favorite_Genre = col_character(),
##   ..   Last_Login = col_date(format = "")
##   .. )
##  - attr(*, "problems")=<externalptr>
```

## 查看資料集行數和列數

```r
rows <- nrow(data)
columns <- ncol(data)

# 設定全域缺失值顯示為 "nan"
options(na.print = "nan")

if (rows < 100 && columns < 20) {
 # 短表資料（行數少於 100 且列數少於 20）查看全量資料資訊
 cat('資料全部內容資訊: \n')
 print(data)
} else {
 # 長表資料查看資料前 6 行訊息
 cat('資料前幾行內容資訊: \n')
 print(head(data))
}
```

```
## 資料前幾行內容資訊:
## # A tibble: 6 x 8
##   User_ID Name     Age Country Subscription_Type Watch_Time_Hours Favorite_Genre
```

```
##     <dbl> <chr>  <dbl> <chr>   <chr>                      <dbl> <chr>
## 1       1 James~    18 France  Premium                     80.3 Drama
## 2       2 John ~    23 USA     Premium                    322.  Sci-Fi
## 3       3 Emma ~    60 UK      Basic                       35.9 Comedy
## 4       4 Emma ~    44 USA     Premium                    262.  Documentary
## 5       5 Jane ~    68 USA     Standard                   909.  Drama
## 6       6 David~    21 USA     Standard                   616.  Romance
## # i 1 more variable: Last_Login <date>
```

## 三、詳細分析過程

### （一）年齡分佈

年齡的描述性統計資訊

```
age_stats <- round(summary(data$Age), 2)

cat('年齡的描述性統計資料：\n')
```

取得年齡的描述性統計訊息，並保留兩位小數

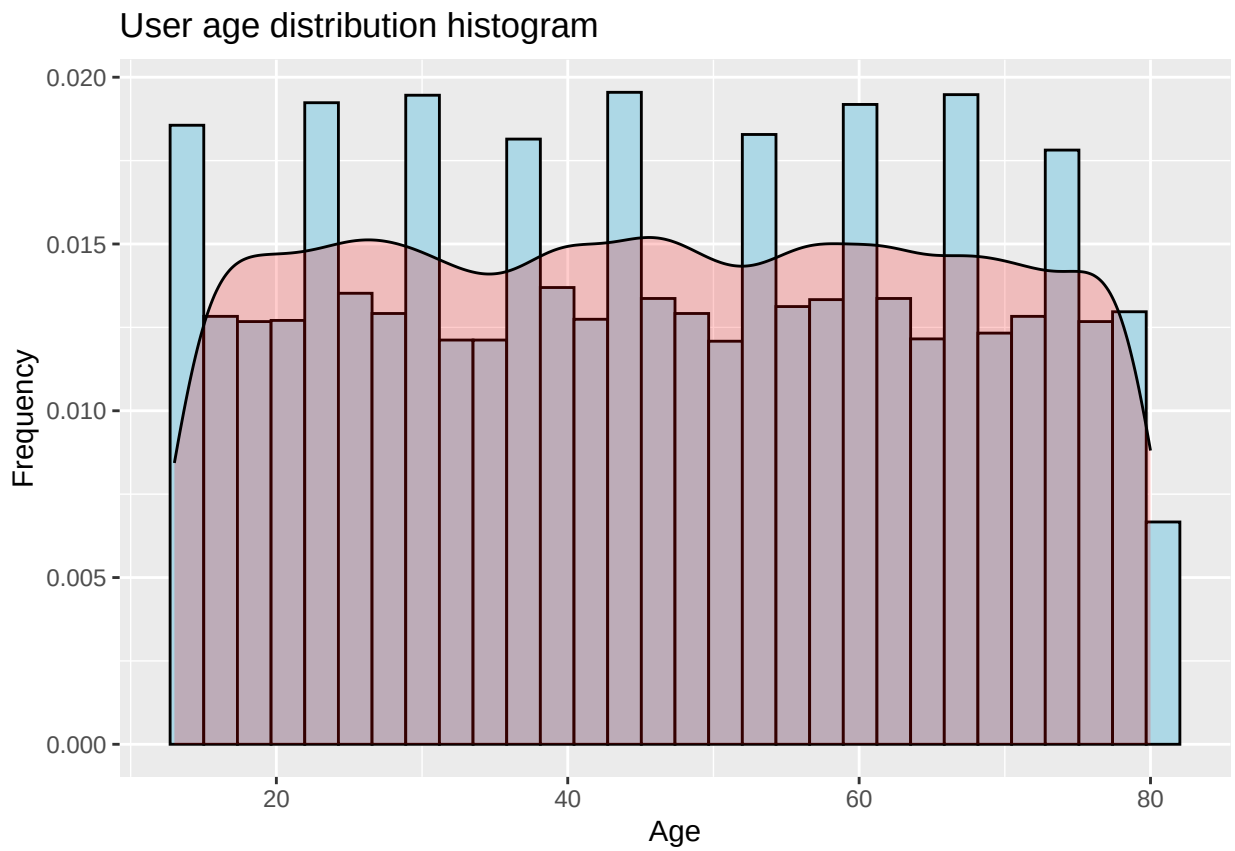
## 年齡的描述性統計資料：

```
print(age_stats)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   29.00   46.00   46.48   63.00   80.00
```

## 繪製年齡直方圖

```
library(ggplot2)
ggplot(data, aes(x = Age)) +
  geom_histogram(bins = 30, aes(y = after_stat(density)), fill = "lightblue", color = "black") +
  geom_density(alpha = 0.2, fill = "red") +
 labs(x = "Age", y = "Frequency", title = "User age distribution histogram")
```




從描述性統計資料我們可以推測到以下內容：

樣本數：此分析的使用者年齡資料樣本量為 25,000 個，樣本量較大，且能較好地反映總體的年齡特徵。

集中趨勢：平均年齡為 46.48 歲，中位數（50% 分位數）為 46.00 歲，二者較為接近，說明年齡分佈可能近似對稱，但仍需結合其他統計量進一步判斷。

離散程度：標準差為 19.59，表示年齡資料相對較為分散。最小值為 13 歲，最大值為 80 歲，年齡跨度較大，這可能表示使用者群體涵蓋了較廣的年齡層。25% 分位數為 29 歲，75% 分位數為 63 歲，代表有 50% 的使用者年齡在 29 - 63 歲這個區間內。

（二）不同國家的使用者數量

```
country_counts <- data.frame(table(data$Country))
colnames(country_counts) <- c("Country", " 使用者數量")

cat('不同國家的使用者數量: \n')
```

不同國家使用者數量統計

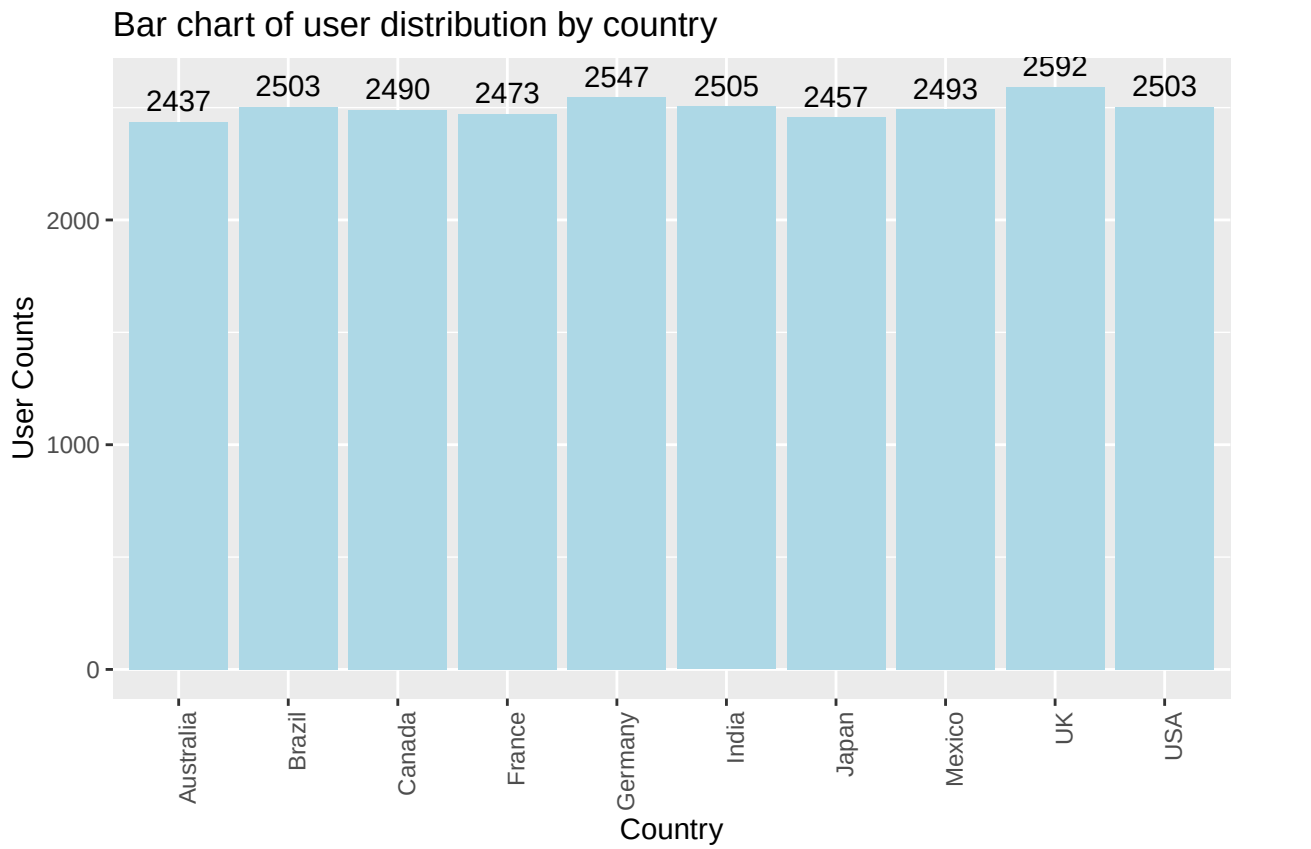
## 不同國家的使用者數量：

```
print(country_counts)
```

```
##       Country 使用者數量
## 1   Australia       2437
## 2      Brazil       2503
## 3      Canada       2490
## 4      France       2473
## 5     Germany       2547
## 6       India       2505
## 7       Japan       2457
## 8      Mexico       2493
## 9          UK       2592
## 10        USA       2503
```

## 繪製長條圖

```
ggplot(country_counts, aes(x = Country, y = 使用者數量)) +
 geom_col(fill="lightblue") +
 geom_text(aes(label = 使用者數量), vjust = -0.5) +
 labs(x = "Country", y = "User Counts", title = "Bar chart of user distribution by country") +
 theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




從不同國家的使用者數量統計結果來看，各國的使用者數量較為接近，差異相對較小。英國的使用者數量最多，為 2592 人，澳洲的使用者數量最少，為 2,437 人。這可能顯示該業務在各國的市場推廣程度、受眾規模等因素較為均衡，沒有出現某個國家的使用者量遠高於或低於其他國家的情況。也有可能是該業務在不同國家採取了相似的營運策略，導致用戶數量分佈比較平均。

（三）不同訂閱類型的平均觀看時間

不同訂閱類型的平均觀看時長統計

```
subscription_avg_time <- aggregate(Watch_Time_Hours ~ Subscription_Type, data = data, FUN = mean)
subscription_avg_time$Watch_Time_Hours <- round(subscription_avg_time$Watch_Time_Hours, 2)

cat('不同訂閱類型的平均觀看時間: \n')
```

計算不同訂閱類型的平均觀看時長，並保留兩位小數

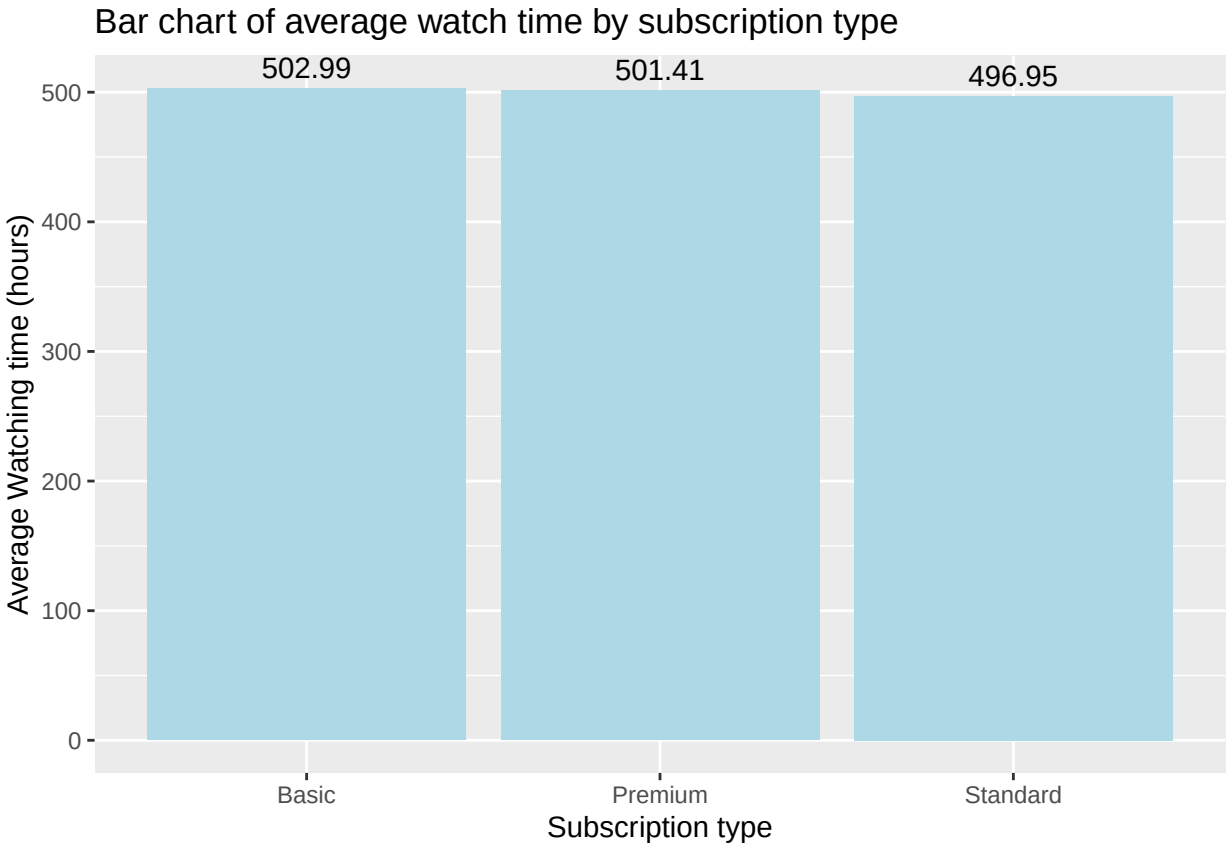
```
## 不同訂閱類型的平均觀看時間:
```

```
print(subscription_avg_time)
```

```
##   Subscription_Type Watch_Time_Hours
## 1             Basic           502.99
## 2           Premium           501.41
## 3          Standard           496.95
```

## 繪製長條圖

```
ggplot(subscription_avg_time, aes(x = Subscription_Type, y = Watch_Time_Hours)) +
 geom_col(fill="lightblue") +
 geom_text(aes(label = Watch_Time_Hours), vjust = -0.5) +
 labs(x = "Subscription type", y = "Average Watching time (hours)", title = "Bar chart of average watch
```




從數據中可以看出，不同訂閱類型的平均觀看時長較為接近。

其中，基礎（Basic）訂閱類型的平均觀看時長最長，為 502.99 小時；標準（Standard）訂閱類型的平均觀看時長最短，為 496.95 小時。

這可能暗示著不同訂閱類型在內容資源、服務品質等方面對用戶觀看時長的影響差異不大，或者用戶的觀看習慣更多地取決於自身需求而非訂閱類型。也有可能是該平台的不同訂閱類型所提供的內容覆蓋範圍相似，導致用戶觀看時長相近。

（四）最受歡迎的三種電影類型

最受歡迎電影類型的使用者數量統計

```
genre_counts <- data.frame(table(data$Favorite_Genre))
colnames(genre_counts) <- c("Favorite_Genre", " 使用者數量")
```

統計不同電影類型的使用者數量

## 找出最受歡迎的三種電影類型

```r
top_3_genres <- head(genre_counts[order(-genre_counts$使用者數量), ], 3)

cat('最受歡迎的三種電影類型: \n')
```

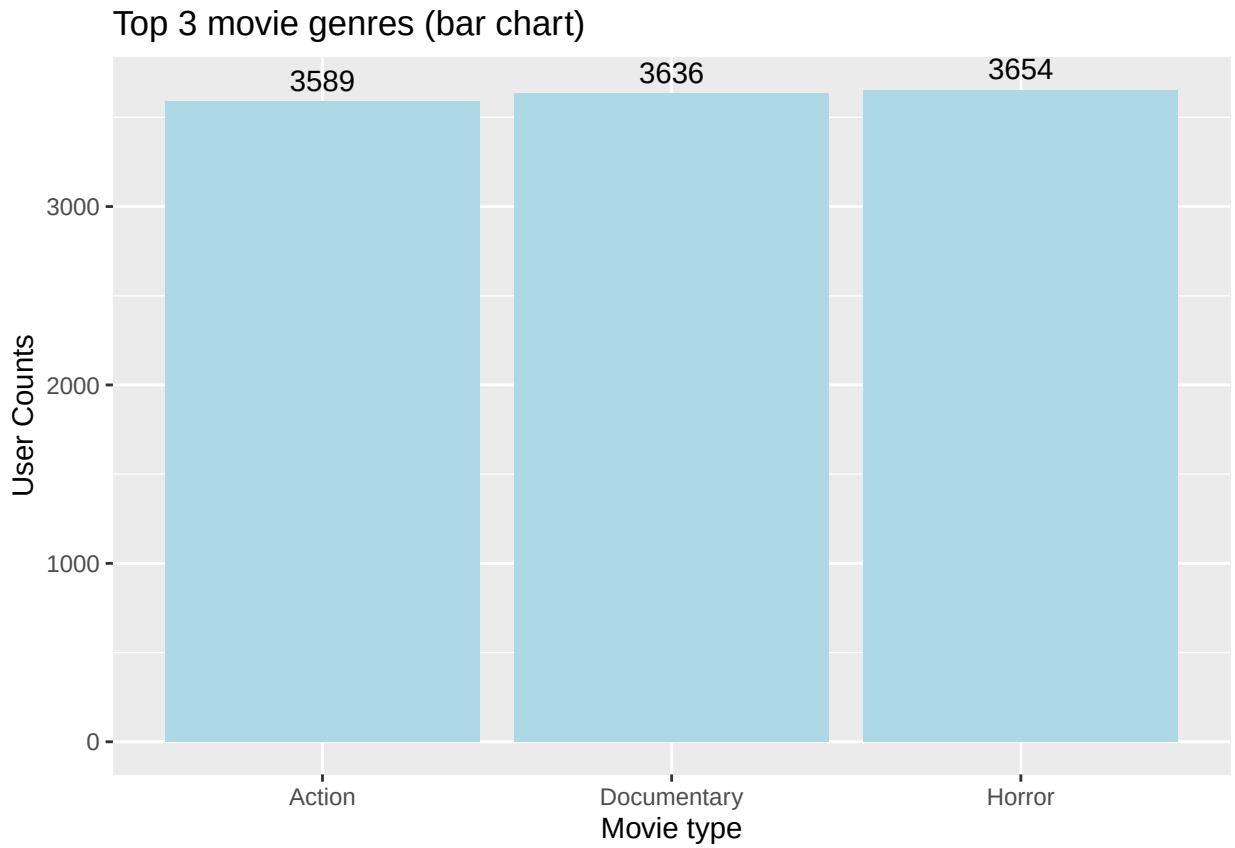
```
## 最受歡迎的三種電影類型:
```

```r
print(top_3_genres)
```

```
##   Favorite_Genre 使用者數量
## 5         Horror       3654
## 3    Documentary       3636
## 1         Action       3589
```

## 繪製長條圖

```r
ggplot(top_3_genres, aes(x = Favorite_Genre, y = 使用者數量)) +
 geom_col(fill="lightblue") +
 geom_text(aes(label = 使用者數量), vjust = -0.5) +
 labs(x = "Movie type", y = "User Counts", title = "Top 3 movie genres (bar chart)")
```




從這些數據可以推測出，在該數據所涉及的範圍內，恐怖（Horror）類型的電影最受用戶喜愛，擁有 3654 名用戶將其列為最喜歡的電影類型。紀錄片（Documentary）和動作（Action）類型也很受歡迎，使用者數量與恐怖類

型非常接近。這可能反映出當前用戶的觀影口味傾向於刺激的恐怖體驗、真實的紀錄片內容或充滿熱情的動作場面。對於電影製作方或影視平台來說，可以根據這些受歡迎的類型來調整內容採購或製作計畫。

（五）每月登入次數的變化趨勢

**1.** 每月登入次數統計

```
data$Last_Login <- as.Date(data$Last_Login)
```

將 **Last__Login** 轉換為日期格式

## 統計每月登入次數

```
monthly_login_counts <- data.frame(table(format(data$Last_Login, "%Y-%m")))
colnames(monthly_login_counts) <- c(" 年月", " 登入次數")
monthly_login_counts <- monthly_login_counts[order(monthly_login_counts$年月), ]

cat('每月登入次數: \n')
```

```
## 每月登入次數:
```

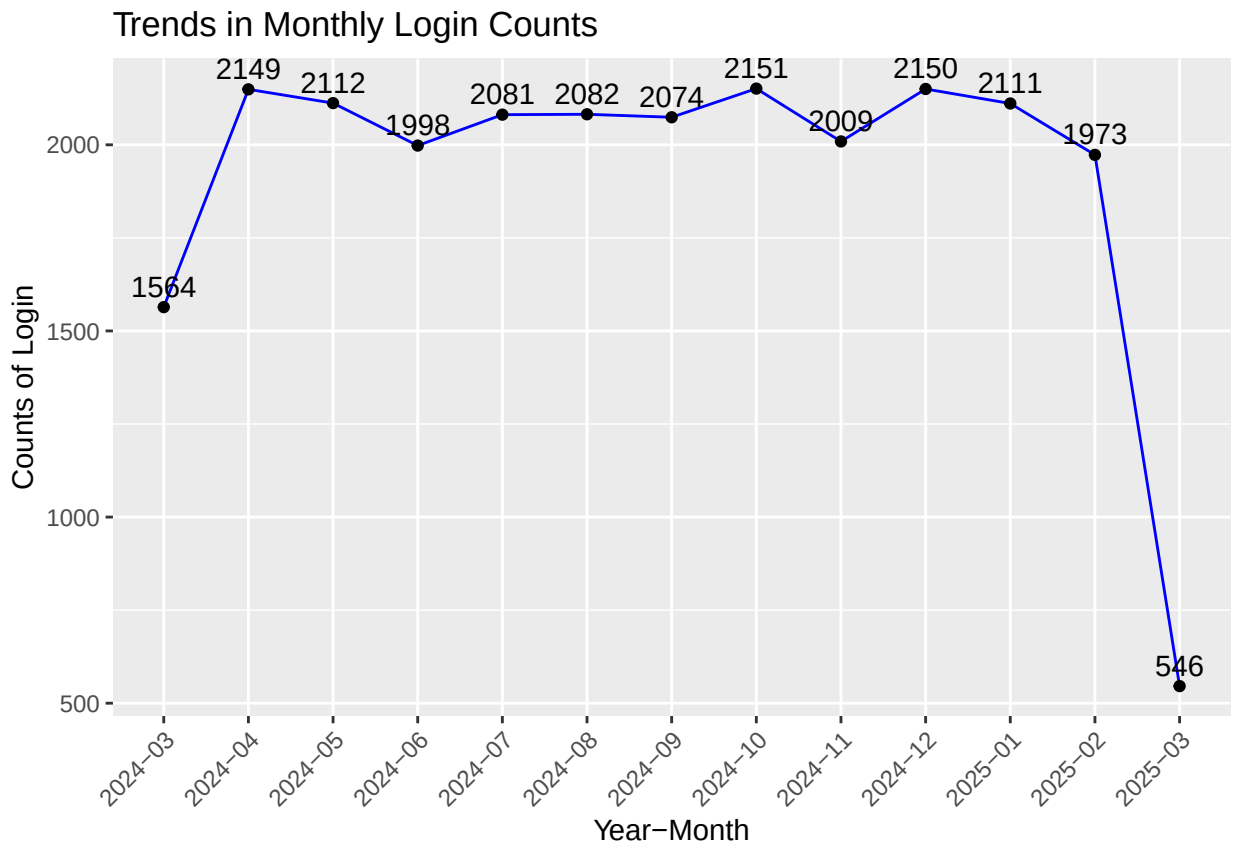
```
print(monthly_login_counts)
```

```
##        年月  登入次數
## 1   2024-03     1564
## 2   2024-04     2149
## 3   2024-05     2112
## 4   2024-06     1998
## 5   2024-07     2081
## 6   2024-08     2082
## 7   2024-09     2074
## 8   2024-10     2151
## 9   2024-11     2009
## 10  2024-12     2150
## 11  2025-01     2111
## 12  2025-02     1973
## 13  2025-03      546
```

## 繪製折線圖

```
ggplot(monthly_login_counts, aes(x = 年月, y = 登入次數, group = 1)) +
 geom_line(color="blue") +
 geom_point() +
```

```
geom_text(aes(label = 登入次數), vjust = -0.5) +
labs(x = "Year-Month", y = "Counts of Login", title = "Trends in Monthly Login Counts") +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




從每月登入次數的數據來看，在 2024 年 3 月至 2025 年 2 月這一統計區間內，登入次數大致在 1973-2151 之間波動，整體處於相對穩定的狀態。然而，2025 年 3 月的登入次數卻出現了顯著下降，僅有 546 次。

需要注意的是，由於本次資料統計的時間跨度僅為 *2024/03/08-2025/03/08*，該月登入次數的大幅下降可能有許多原因。

一方面，基於現有資料無法排除平台在 2025 年 3 月進行系統維護或技術故障的可能性，這些因素可能導致使用者登入受阻，進而使登入次數降低。

另一方面，在這個時間段內，或許有競爭對手推出了更具吸引力的服務，不過鑑於數據的局限性，目前難以明確競爭對手的具體動作及其對本平台用戶的影響程度。

## 四、結論

透過 Netflix 在 2024/03/08-2025/03/08 期間用戶資料的詳細分析，我們對使用者的特徵和行為有了較為全面的了解：

-年齡分佈：使用者群體涵蓋了較廣的年齡層，平均年齡為 46.48 歲，年齡跨度從 13 歲至 80 歲，這表明平台吸引了不同年齡層的用戶。

-國家分佈：不同國家的使用者數量分佈較為均衡，英國使用者數量最多（2,592 人），澳洲使用者數量最少（2,437 人），各國家間使用者數量差異相對較小，反映出平台在不同國家市場的發展較為穩定。

-訂閱類型：不同訂閱類型（基礎、進階、標準）的平均觀看時長差異不大，基礎訂閱類型平均觀看時長最長（502.99 小時），標準訂閱類型最短（496.95 小時），這可能意味著不同訂閱類型提供的內容對用戶觀看時長的影響有限，用戶觀看習慣可能更多取決於個人喜好。

-電影類型偏好：恐怖、紀錄片和動作類型電影最受用戶歡迎，選擇這三種類型的用戶數量分別為 3654 人、3636 人、3589 人。這為平台的內容採購和製作方向提供了參考。

-登入次數波動：每月登入次數在 2025 年 3 月大幅下降，儘管可能存在平台維護、技術故障或競爭等因素，但由於資料時間範圍的限制，還需要進一步收集更多資料進行深入分析，以便更準確地判斷導致登入次數下降的原因。

## 五、策略調整

基於以上分析結果，可以考慮採取以下針對性的策略調整:

-內容推薦方面: 根據用戶對電影類型的偏好，加大恐怖、紀錄片和動作類影片的推薦比重，提高用戶發現感興趣內容的概率，從而提升用戶體驗。

-市場推廣方面: 鑒於不同國家用戶數量分佈均衡，平台可進一步鞏固現有市場，同時針對用戶數量相對較少的國家，制定更具針對性的推廣策略，挖掘潛在用戶群體，進一步拓展市場份額。

-運營方面: 平台需要密切關注每月登錄次數的波動情況，尤其是出現異常下降的月份，及時排查技術問題，確保平台的穩定運行。此外，持續關注競爭對手的動態，不斷優化自身服務，增強平台的核心競爭力，以提高用戶的滿意度和忠誠度，促使用戶更長期、更活躍地使用平台。

*Open data from Kaggle Netflix Users Database*