國立臺灣大學

程式設計與資料科學導論 112-1 Final Project

1992-2020 新興韓團火熱程度關聯分析

Group 17

王學謙、李采蓉、賴郁升

B11103019、B08207042、R12228019

supervised by

謝舒凱

December 27, 2023

一 、簡介

當前的 K-pop 現象已成為全球文化的重要代表之一，不僅僅是大型知名團體獲得了廣泛的關注和歡迎，同時也涌現出眾多中小型團體，這使得市場競爭日益激烈。在這樣競爭激烈的環境中，如何在眾多風格迥異、表現亮眼的團體中脫穎而出，成為了眾多從業者不斷探討和試圖解答的難題。

為了深入研究這個問題，本研究專案特別從 Kaggle 上收集了相關的資料集，同時也結合了透過爬蟲技術從 YouTube 獲取的大量數據，包括但不限於觀看次數、點讚數、留言數等。這些數據的探索與分析，旨在尋找、解析和歸納出韓團成功的關鍵要素。透過對數據的細致觀察和深入分析，我們試圖揭示出背後可能存在的模式、豁勢和關聯，進而為行業決策者提供更多洞見和啟發。這項研究的目的是不僅僅為了探索韓團成功的秘訣，更是為了在這個繁榮蓬勃的行業中，為未來的團體提供寶貴的策略和建議。

二 、資料選用以及資料集的限制

2.1 主要資料集

採用資料集 K-Pop Database (1992-2020)，根據 Kaggle 上的說明，此份資料整理自 K-Pop Database 這個網站。

下列是此次分析探討的三個檔案及其欄位表：

男性組成韓國流行樂團體資料

Name	Short	Debut	Company	Members	Orig. Memb.	Fanclub Name	Active

此份資料共紀錄 147 個男性組成的韓國流行樂團體。

女性組成韓國流行樂團體資料

Name	Short	Debut	Company	Members	Orig. Memb.	Fanclub Name	Active

此份資料共紀錄 152 個女性組成的韓國流行樂團體。

韓國流行樂影片資料

Date	Artist	Song Name	Korean Name	Director	Video	Type	Release

此份資料共紀錄 3443 筆韓國流行樂影片。

2.2 次要資料集

利用 YouTube Data API，從 *kpop_music_videos.csv* 提供的 `Video` 中的網址，爬取其 `Views`、`Likes`、`Comments` 數量，輸出成 *youtube.csv*。我們注意到有些影片的點閱率、點讚數、評論區等被鎖定，因此我們刪除了這些資料。

另外，我們分配一人查找 99 個團的成員出道年齡，最後整合成每個團體出道時的平均年齡，並將其命名為 `Average_age`。

2.3 資料集的限制

此資料集在 kaggle 上的信用分數僅 7.06 分，代表

三 、研究方法

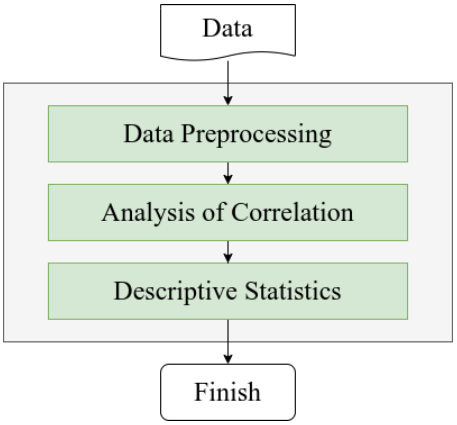

圖 1: 整體流程

3.1 資料前處理

在合併 *kpop_idols_boy_groups.csv* 和 *kpop_idols_girl_groups.csv* 的過程中，我們新增了一個 `Gender` 欄位，並分別標記為 M 和 F。接著，我們去除了多餘的欄位，保留了 `Name`、`Debut`、`Company`、`Orig. Member` 和 `Gender` 欄位。同時，利用 `pd.factorize` 方法將 `Company` 轉換為 `Company Code`，並將 `Name` 設置為索引。我們也將次要資料集的 `Average_age` 加入了主要資料集。最後的資料集樣貌如下表所示：

Name	Debut	Company	Orig. Memb.	Average_age	Gender	Company Code

針對次要資料集 *youtube.csv*，在未處理前其內容如下：

Name	Date	Likes	Views	Comments

我們的計算方法是先算出每首歌的平均點閱率、讚數和留言數，然後將各欄位數值相加取平均。如果某團體沒有任何一首歌被收錄進此資料集，則移除該團體。經過處理後，我們得到了總共剩下 **239** 個韓國流行樂團體的資料。

3.2 初步分析

根據男團出道年齡的分析，資料顯示了出道年份為 1995 年的男團平均出道年齡為 14.00 歲，為最年輕的出道年份。而在男團中，出道年份為 2010 年的團體平均出道年齡為 22.00 歲，為最高的出道年份。男團出道年齡的樣本數據呈現出以下統計特徵：其算術平均數為 20.959，中位數為

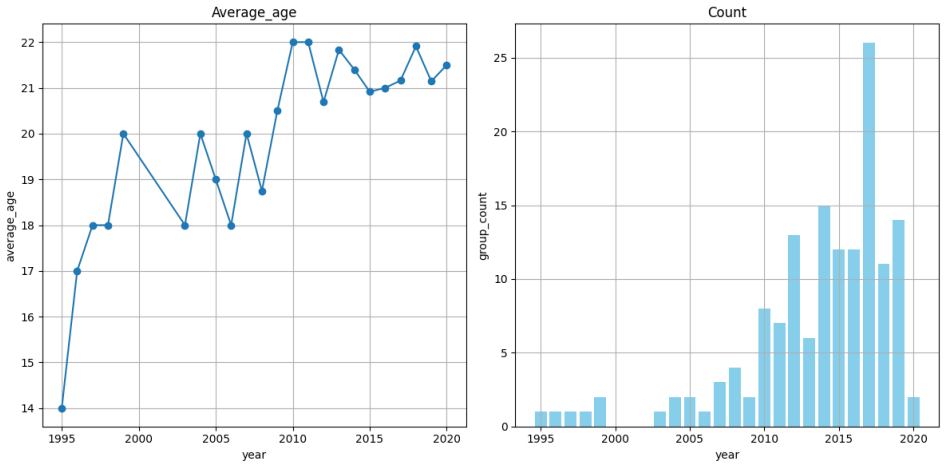

圖 2: Variation Trends in Average Debut Age and Year(Boy)

21.0，而出現最頻繁的眾數為 20.0。此外，男團出道年齡樣本數據的數值範圍為 13.0，四分位數 Q1 為 20.0，Q2（中位數）為 21.0，Q3 為 22.0。針對男團出道年齡樣本數據的統計量，其標準差為 2.347，變異數為 5.509，偏度為-0.390，峰度為 0.798。

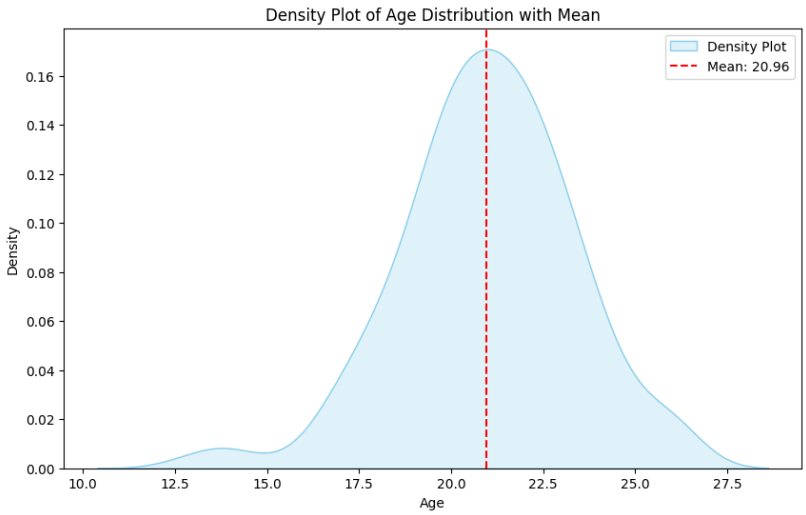

圖 3: Density Plot of Age Distribution with Mean(Boy)

女團出道平均年齡與年份的變化豁勢如下：出道年份最小的為 1997 年，其平均年齡為 16.16 歲；相對地，出道年份最大的為 2006 年，其平均年齡為 23.00 歲。

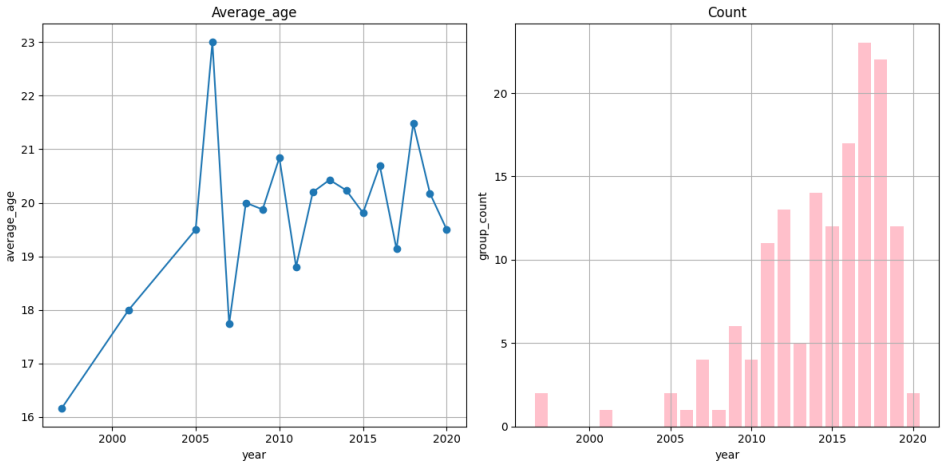

圖 4: Variation Trends in Average Debut Age and Year(Girl)

這些數據來自女團出道年齡樣本的分析，資料顯示出以下統計特徵：其算術平均數為 20.024，中位數為 20.0，最頻繁出現的眾數為 20.0。此外，女團出道年齡樣本數據的數值範圍為 15.0，四分位數 Q1 為 18.513，Q2（中位數）為 20.0，Q3 為 21.0。另外，針對女團出道年齡樣本數據的統計分析，其標準差為 2.417，變異數為 5.841，偏度為 0.302，峰度為 0.496。

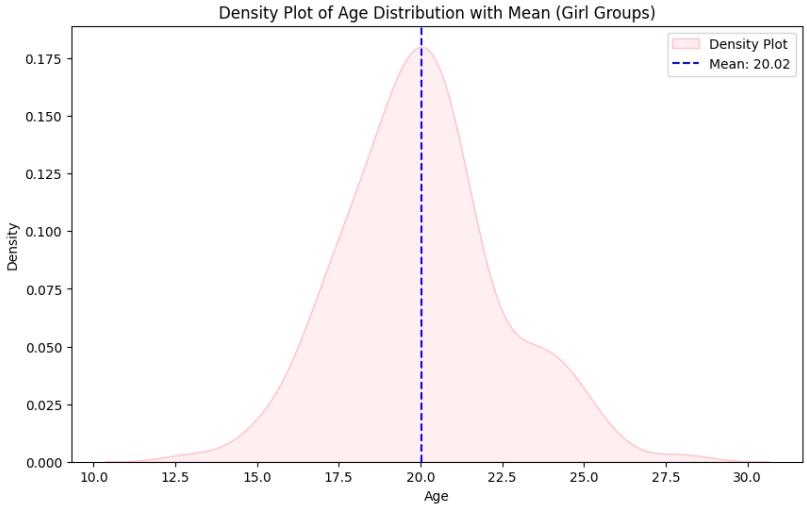

圖 5: Density Plot of Age Distribution with Mean(Girl)

圖 (6) 為出道團體人數與年份之間的散佈圖，團體人數很長一段時間都落在平均之下。隨著韓團數量的增加，人數還是集中在平均上下，但也出現一些人數更多的團體，此情況在男團更為明顯。

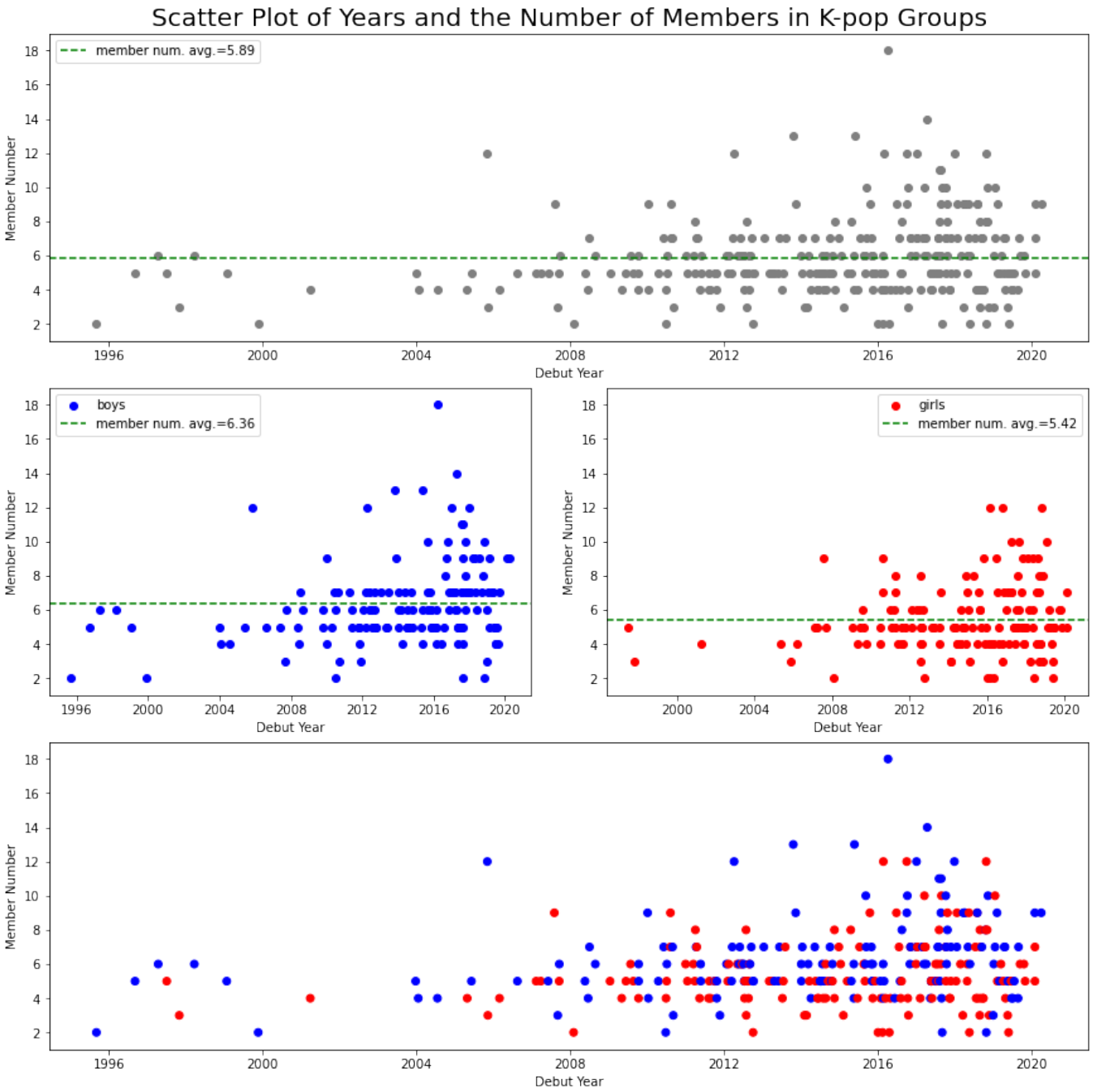

圖 6: Scatter Plot of Years and the Number of Members in K-pop Groups

圖 (7) 為出道團體平均年齡與年份之間的散佈圖，男團的平均年齡較女團略高，但綜合而言皆落在 20 歲的區間中。隨著年份，韓團的數量增加外，也出現更廣的出道平均年齡範圍。

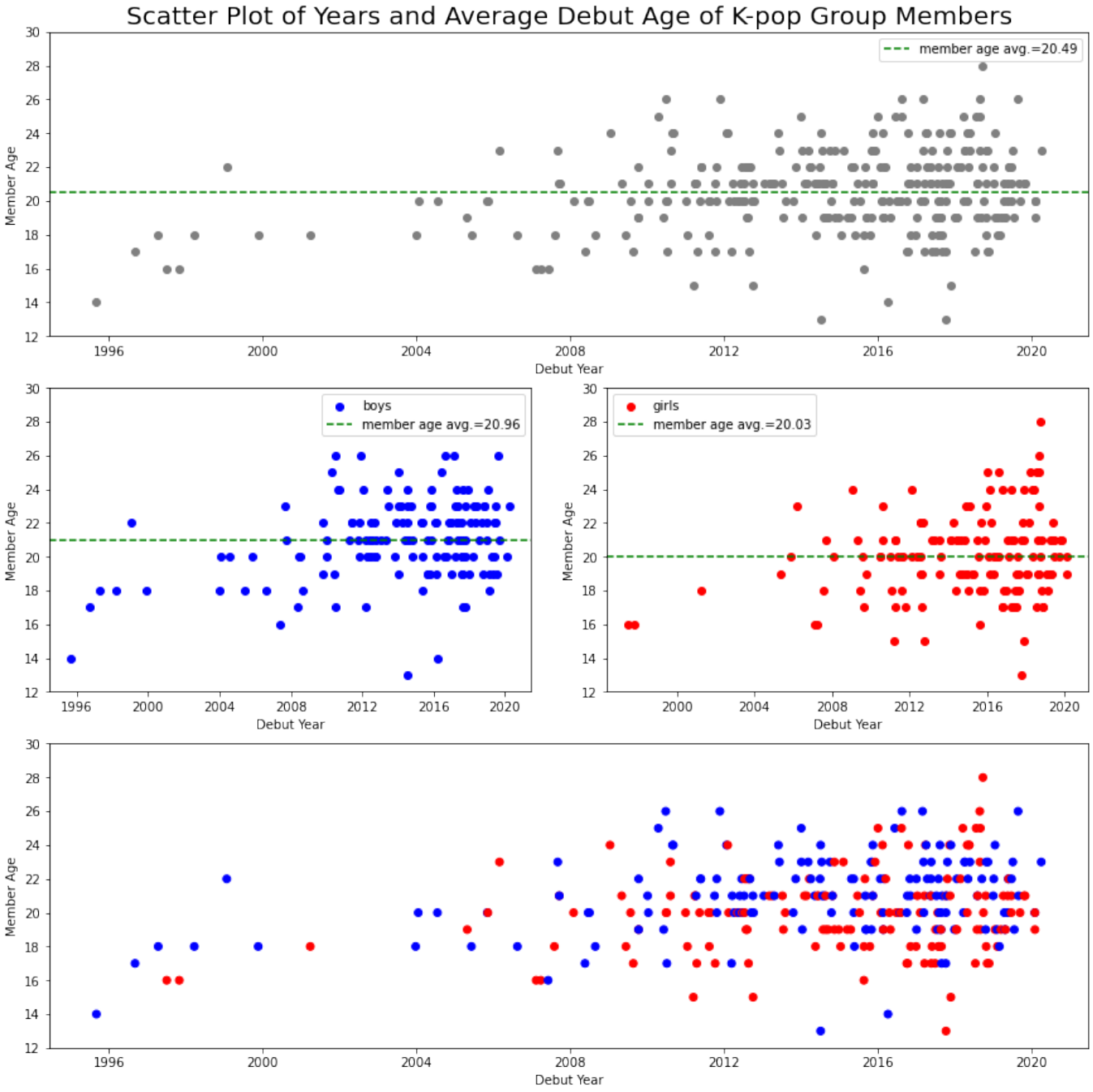

圖 7: Scatter Plot of Years and Average Debut Age of K-pop Groups

接著我們對團體的出道平均年齡和團體人數畫散佈圖 (8)，在不同性別間呈現不太一樣的集中豁勢，女性會較男性更為分散，反應在團體人數跟出道年齡上女團較為多元。

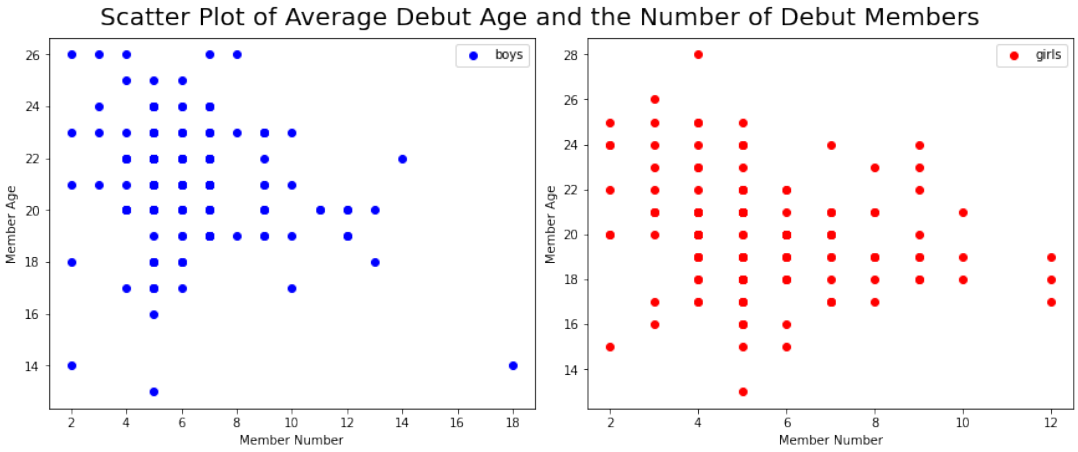

圖 8: Scatter Plot of Average Debut Age and the Number of Members of K-pop Groups

3.3 多元迴歸分析

針對我們所選取的五個變項做相關矩陣，值得注意的是 `Company` 和 `Gender` 間的相關為-0.5，`Company` 和 `Debut Year` 間的相關為 0.52，皆為中度相關。

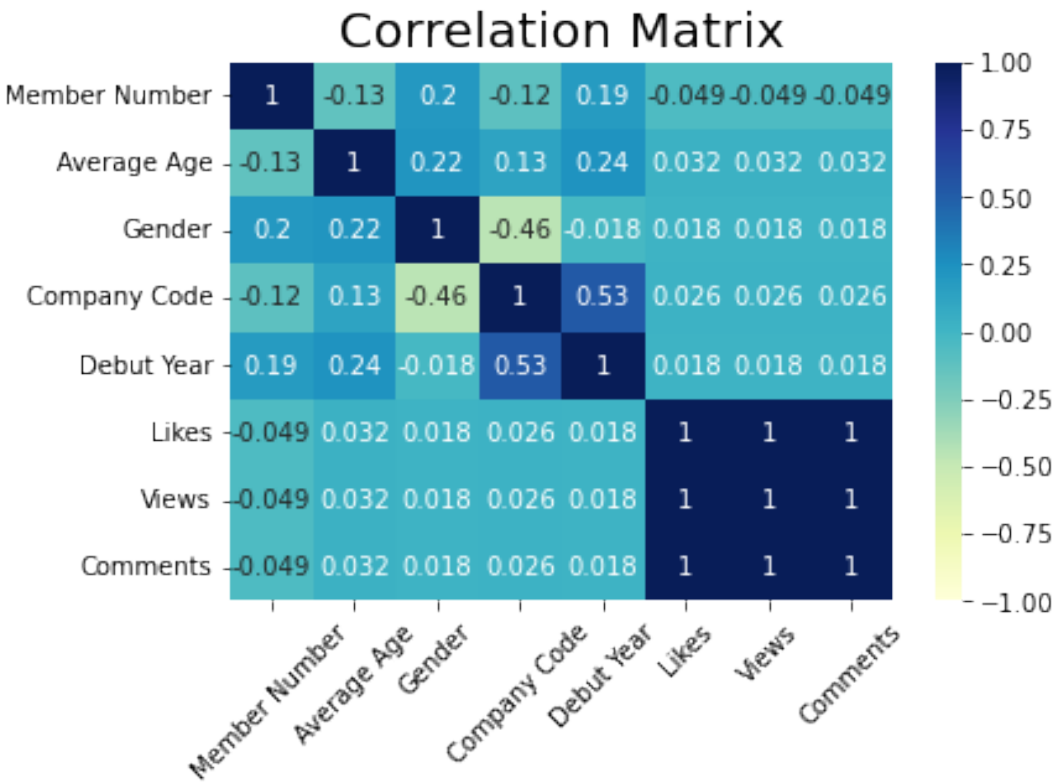

圖 9: Correlation Matrix

縱使相關係數皆成低度相關，我們依舊嘗試套進 Linear Regression 中跑多元迴歸分析，得到下表：

Y	**X**	**Accuracy**
Likes	Member Number, Average Age, Gender, Company, Debut Year	0.4734814692184175%
Views	Member Number, Average Age, Gender, Company, Debut Year	0.4734814691657152%
Comments	Member Number, Average Age, Gender, Company, Debut Year	0.4734814692189837%

表 1: Linear Regression Result

四 、結論

本研究深入探討了近年來 K-pop 團體成功的關鍵要素，從數據分析中發現了一系列有趣的豁勢和特點。首先，男團與女團的出道年紀差異明顯，大多數男團的成員出道年紀落在 21 至 22 歲區間，而女團則以 20 歲為主，但女團成員年齡的差異較男團更為顯著。此外，在團體成員年齡組合方面，女團呈現較為集中的分布，而男團則較為分散。這可能與訓練時長、經紀公司以及男性入營與團體活動等因素有關。與公司相關的因素也受到了關注，發現公司與性別、公司與出道年份存在較高的相關性。這或許表明小公司多半僅推出單一團體，若未取得成功可能後續發展受限。而 2008 年後韓團數量增加的豁勢可能與當時韓國政府所推出的「文化藝術振興法」有關，韓國的影視、音樂、文化從這之後被推廣到全世界，許多小型公司也趁勝追擊推出團體加入競爭。

然而，本研究在多元迴歸分析方面遭遇了困難，可能因為樣本數較少且資料可靠度不高。這促使未來研究可考慮加強特徵選擇、增加 youtube 影片樣本數，以提高研究的準確性和全面性。Kaggle 所提供的資料集或許無法完整呈現一個團體在 2020 年前所有的 youtube 影片，這也是未來研究可加強的方向。

總的來說，這些發現為研究 K-pop 團體成功的因素提供了參考，並為未來相關研究開啟了新的方向與可能性。

五 、工作分配

下表為第十七組的工作分配：

成員名稱	工作項目
王學謙	資料整理、爬蟲程式撰寫、口頭報告發表
李采蓉	資料整理、相關分析、期末報告發表、介紹網頁撰寫、書面報告撰寫
賴郁升	資料整理、描述統計、期末報告發表