圖片：Childish Gambino

Agenda

Process and Methods

一、論文摘要

二、資料變數解釋

三、什麼是機器學習？

四、Lasso Regression (套索算法) 與 SVM (支持向量機)

五、SVR Regression (支持向量機迴歸)

六、2SLS (兩階段最小平方估計量)

七、結論

圖片：Ed Sheeran

論文摘要

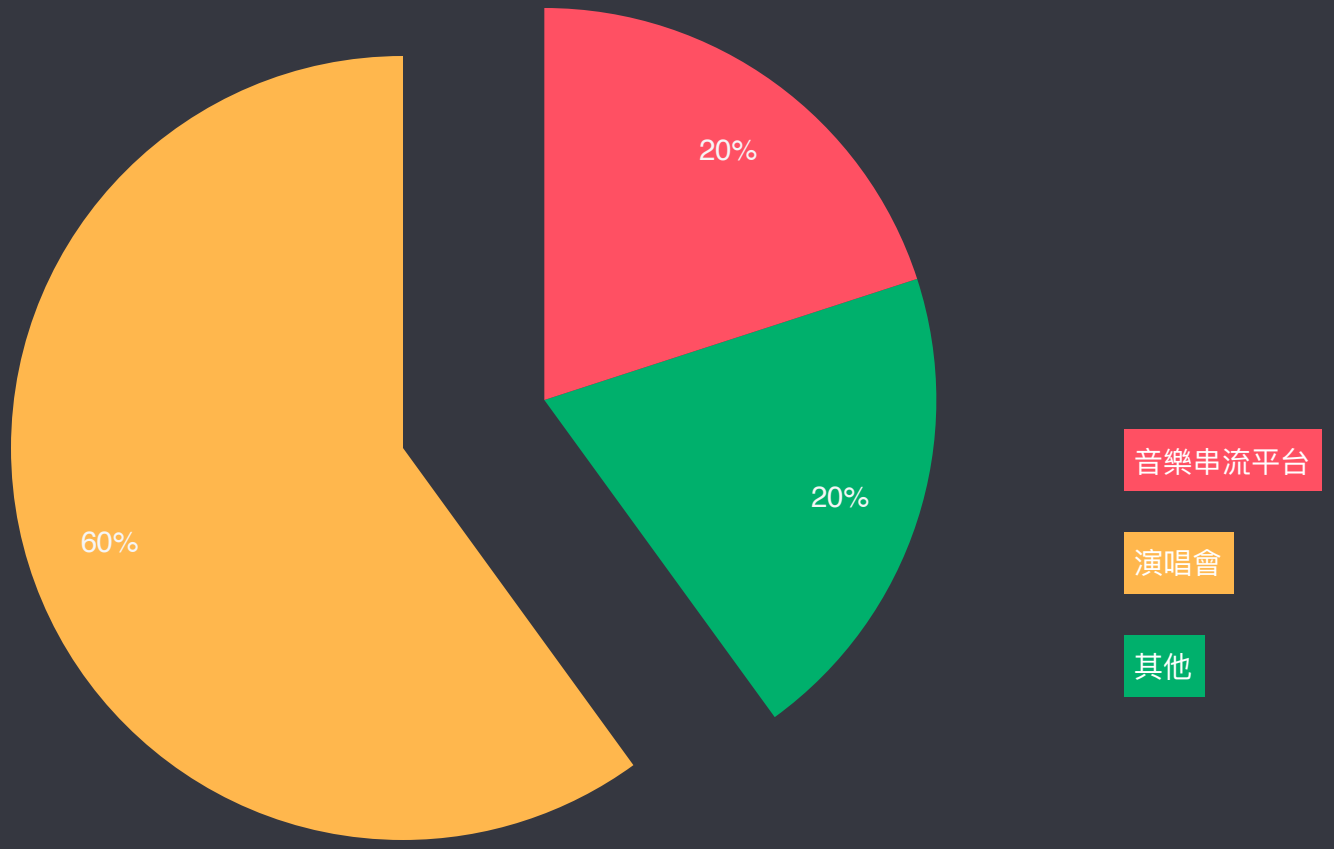
音樂人收入來源

演唱會是音樂人最主要的收入

目前舉辦演唱會所獲得的收入佔音樂人收入約**60%**。
音樂串流平台則佔了**20%**，這邊留給嘉羽講。

其他收入：

- 衣服、及其週邊商品。
- 實體唱片。
- VR體驗。



什麼人會去聽演唱會？

族群略分為三群

有錢有閒，略有印象

人數不多，但通常會坐在最貴的位置。
也包含陪朋友去的。

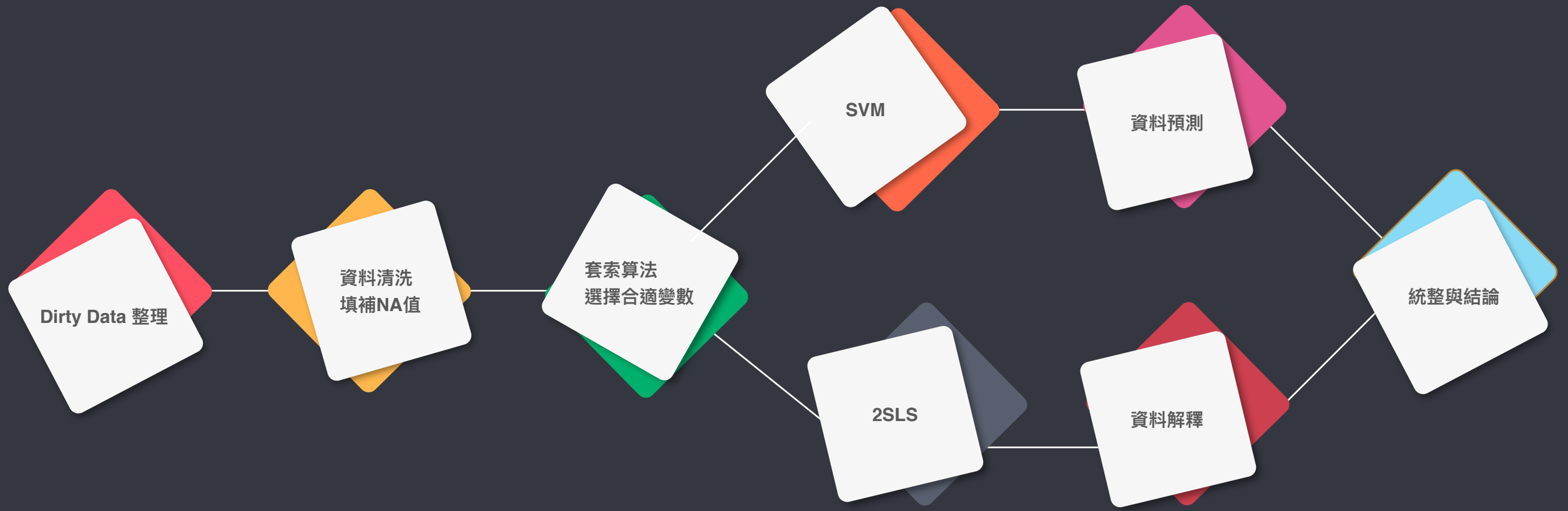
該音樂人的鐵粉

形塑一種鐵粉一定要去支持演唱會，最好多刷幾次，每場都到，希望給音樂人留下印象。大多數支持演唱會的皆為此種族存。

想體驗大型演唱會

人數最少，從未體驗過大型演唱會。

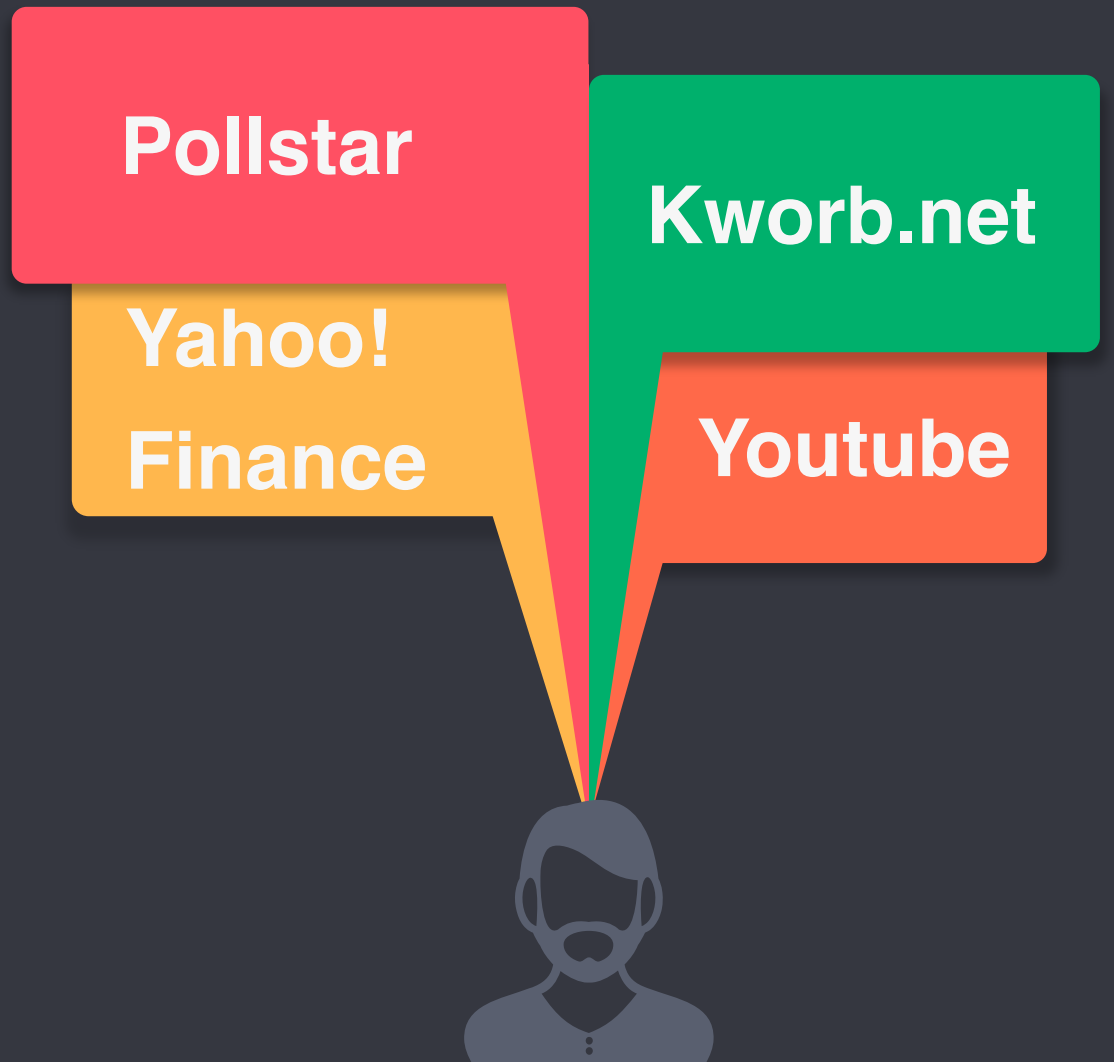
論文研究方法流程


Dirty Data 整理
資料清洗
填補NA值
套索算法
選擇合適變數
SVM
2SLS
資料預測
資料解釋
統整與結論

圖片：Imagine Dragon

資料變數解釋

資料來源



資料變數有哪些？

包含48個變數

- Date：演唱會舉辦的日期。
- Shows：該演唱會辦了幾場表演。
- Year：演唱會舉辦的年份。
- Month：演唱會舉辦的月份。
- Artist：音樂人名字。
- Venue：演唱會舉辦地點。
- City：演唱會舉辦的城市及國家。
- Promoters：演唱會之策展人。
- TicketSold：該場演唱會賣出幾張票。
- Capacity：該場地能容納之人數。
- Percentage：售票率。
- SoldOut：是否完售，為虛擬變數。
- US_Gross：演唱會票房總收益。
- min_price：該場演唱會之最低票價。
- max_price：該場演唱會之最高票價。
- avg_price：該場演唱會之平均票價。
- US_Gross.TicketSold：票房總收益除以總售出票數，為每票收益。
- Supporting：是否有共演團體（暖場表演）。
- Global_Concert：是否有登上百大演唱會排行榜，為虛擬變數。
- Global_Concert_Times：登榜次數。
- InHouse_Promotion：場地擁有者是否同時為策展人。
- no_promoter：策展人的數量。
- State：演唱會舉辦之國家。
- US：是否舉辦在美國，為虛擬變數。
- Canada：是否舉辦在加拿大，為虛擬變數。
- Europe：是否舉辦在歐洲，為虛擬變數。
- UK：是否舉辦在英國，為虛擬變數。
- SouthUS：是否舉辦在美國南部，為虛擬變數。
- Oceania：是否舉辦在大洋洲，為虛擬變數。
- Date_Whichday：舉辦演唱會的日期為哪一天。
- Live.Nation：策展人是否為Live Nation，為虛擬變數。
- Holiday：舉辦演唱會該天是否為假期，不包含週末，為虛擬變數。
- Open.x：Alphabet 當天開盤價格。
- High.x： Alphabet 當天漲停價格。
- Low.x： Alphabet 當年跌停價格。
- Adj.Close.x： Alphabet 當天收盤價格。

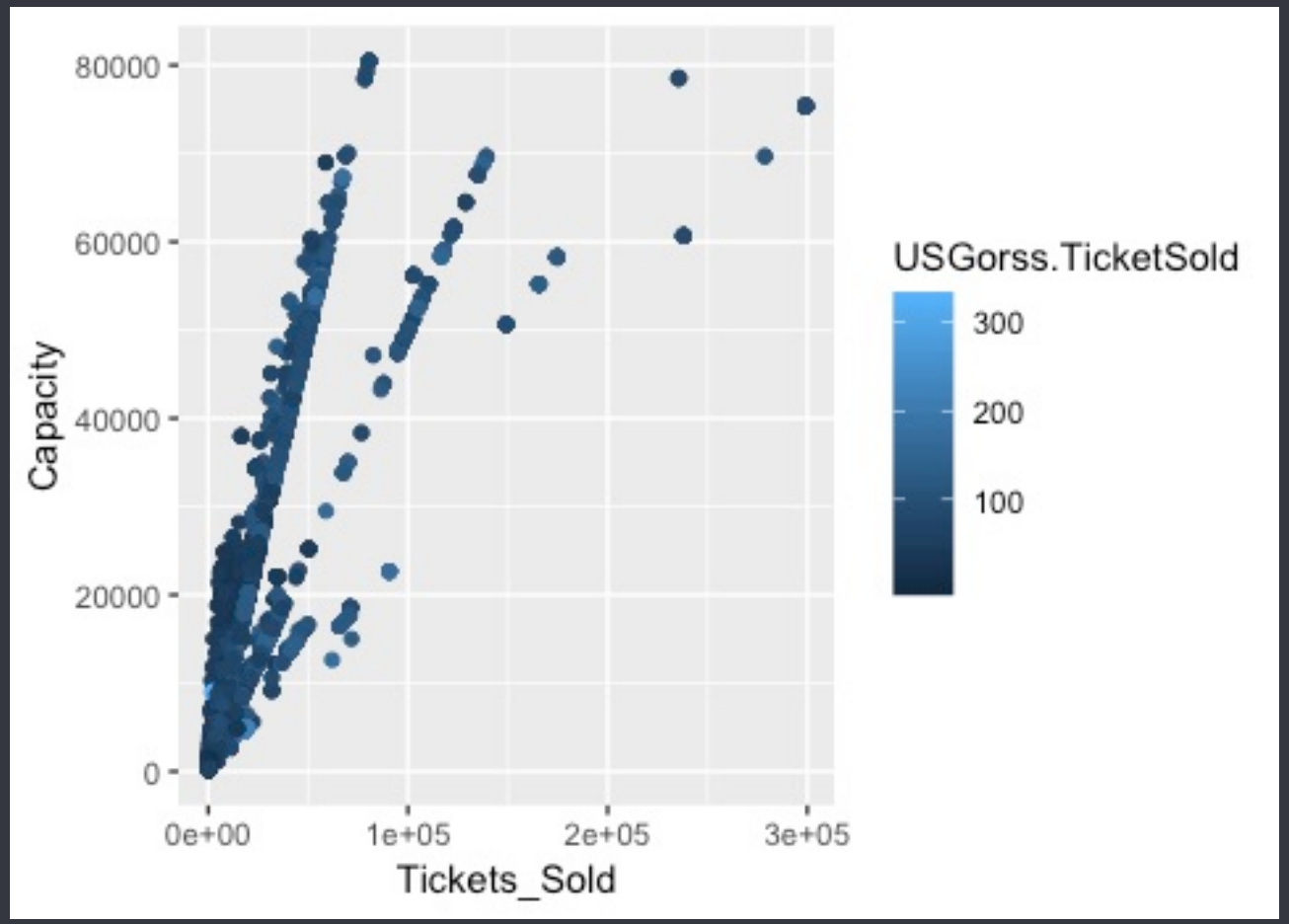
資料變數有哪些？

包含48個變數

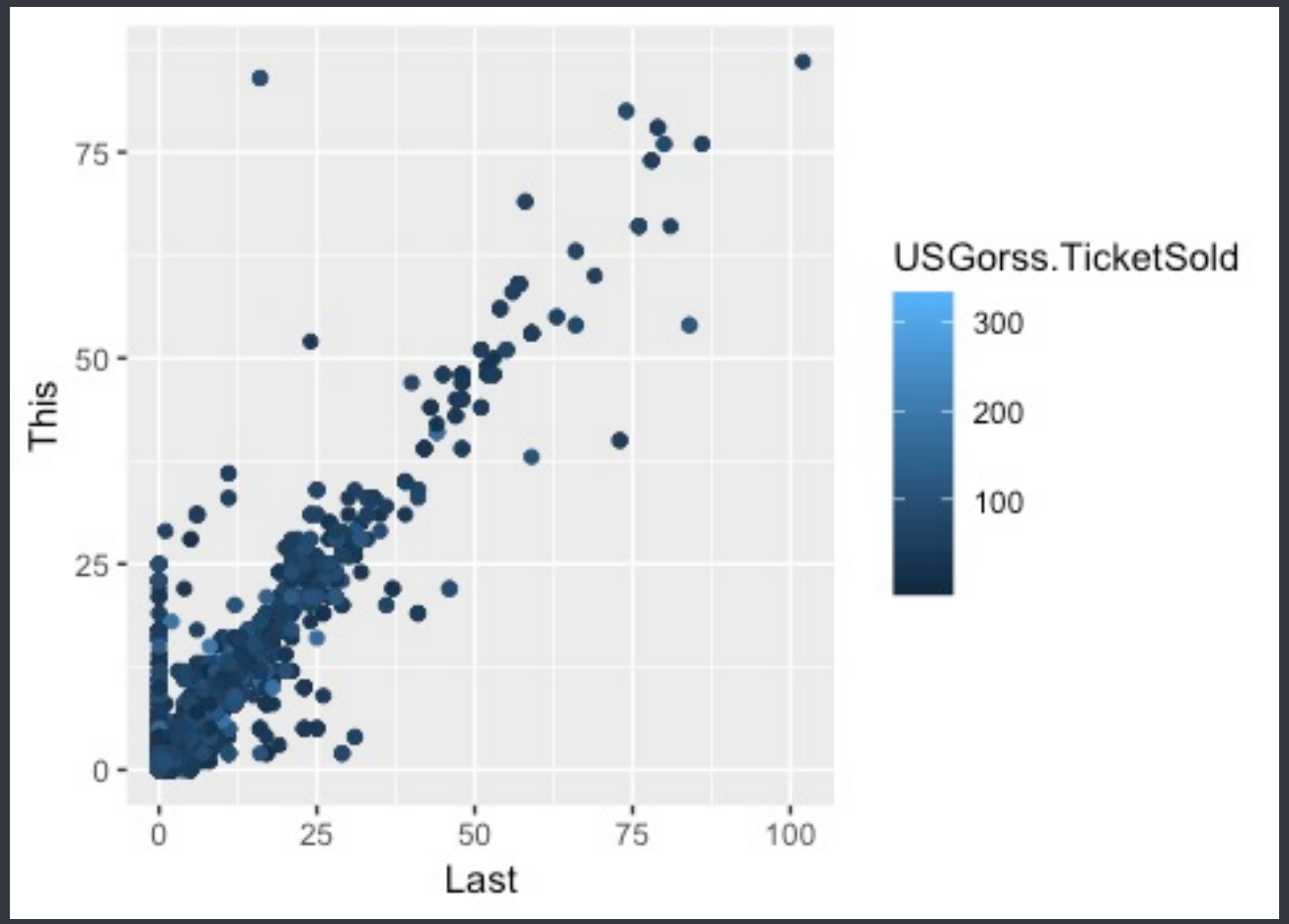
- Adj.Close.x：Alphabet 當天收盤調整後價格。
- Volume.x：Alphabet 當天成交量。
- Open.y：Live Nation 當天開盤價格。
- High.y：Live Nation 當天漲停價格。
- Low.y：Live Nation 當天跌停價格。
- Close.y：Live Nation 當天收盤價格。
- Adj.Close.y：Live Nation 當天收盤調整後價格。
- Volume.y：Live Nation 當天成交量。
- Last：舉辦演唱會前，Youtube歌曲瀏覽量。
- This：舉辦演唱會後，Youtube歌曲瀏覽量。
- Change：舉辦演唱會前後，Youtube歌曲瀏覽量之變化量。
- new_song：該場演唱會前是否發布新歌。

https://github.com/ritalinyutzu/Thesis

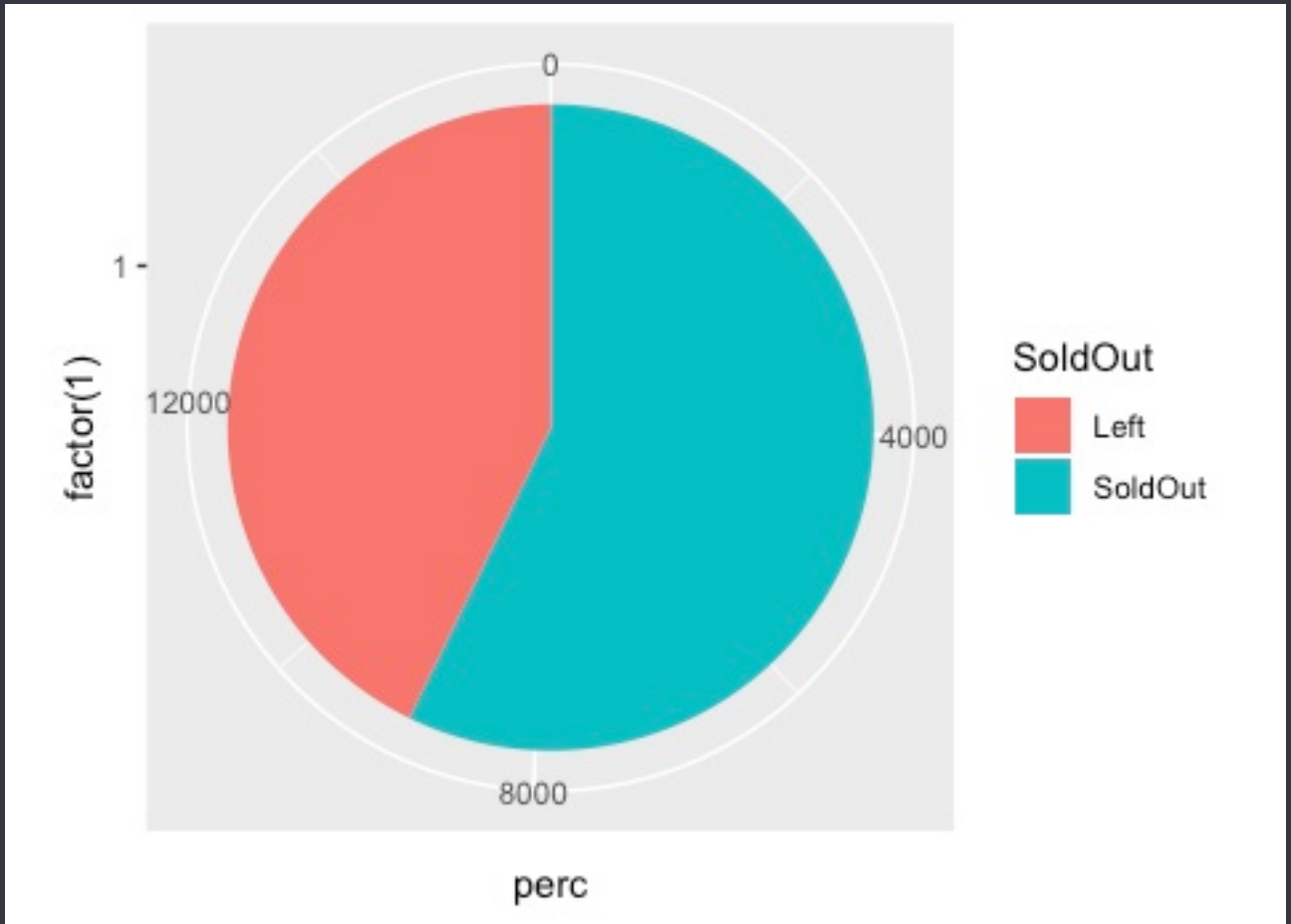
售票率與場地可納人數對每票收益的散佈圖



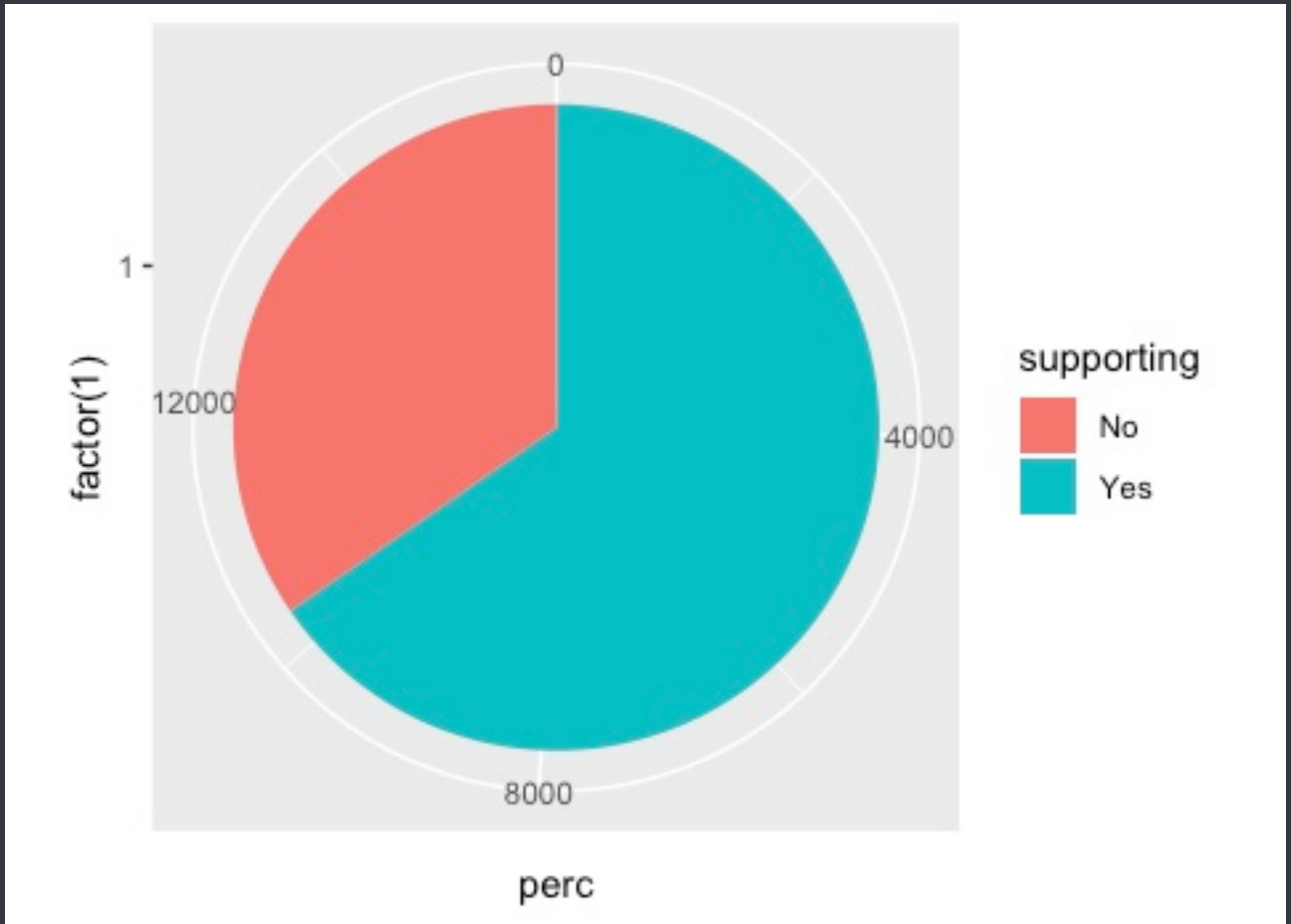
上週點擊率、本週點擊率對每票收益的散佈圖



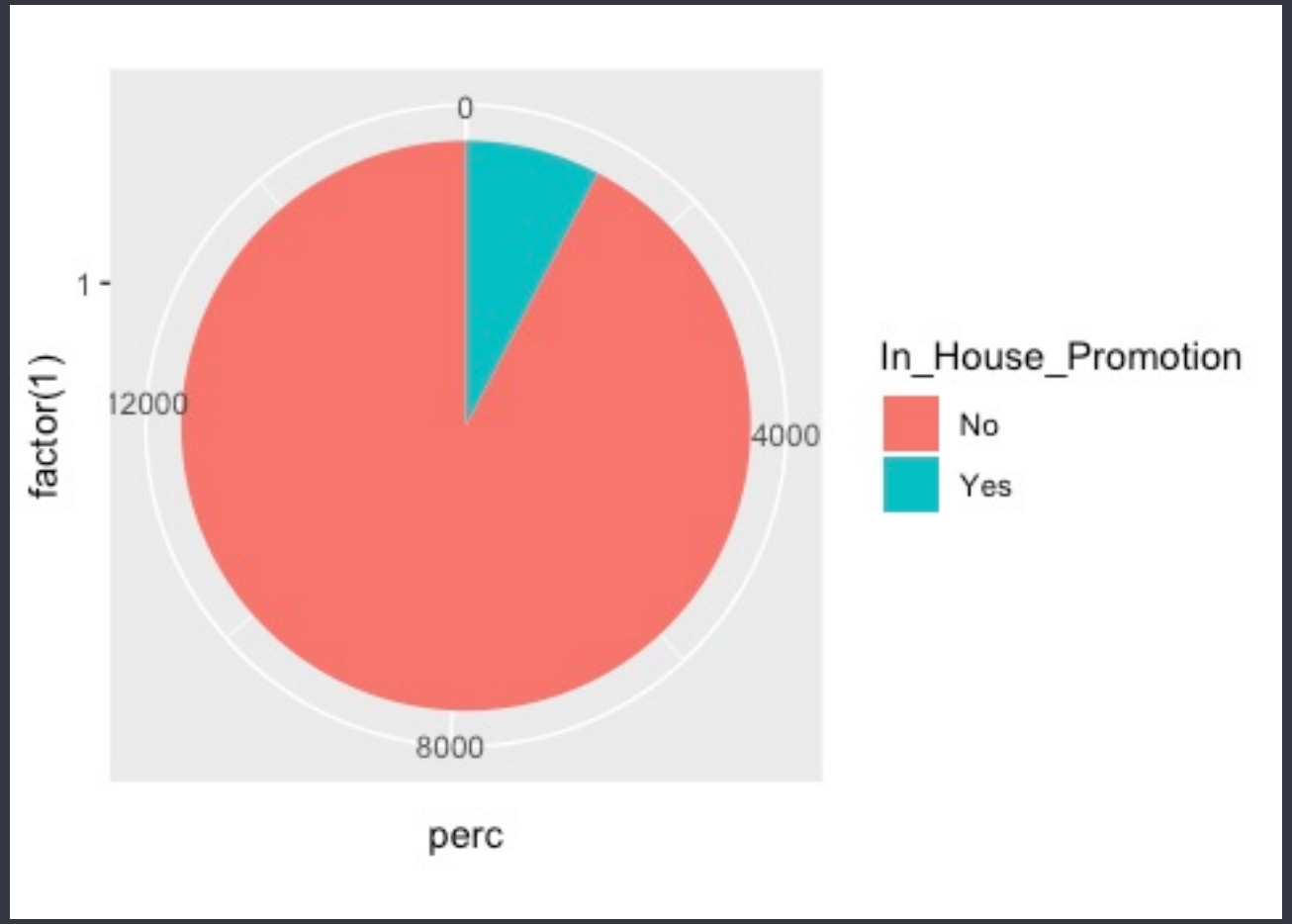
本資料完售率表現



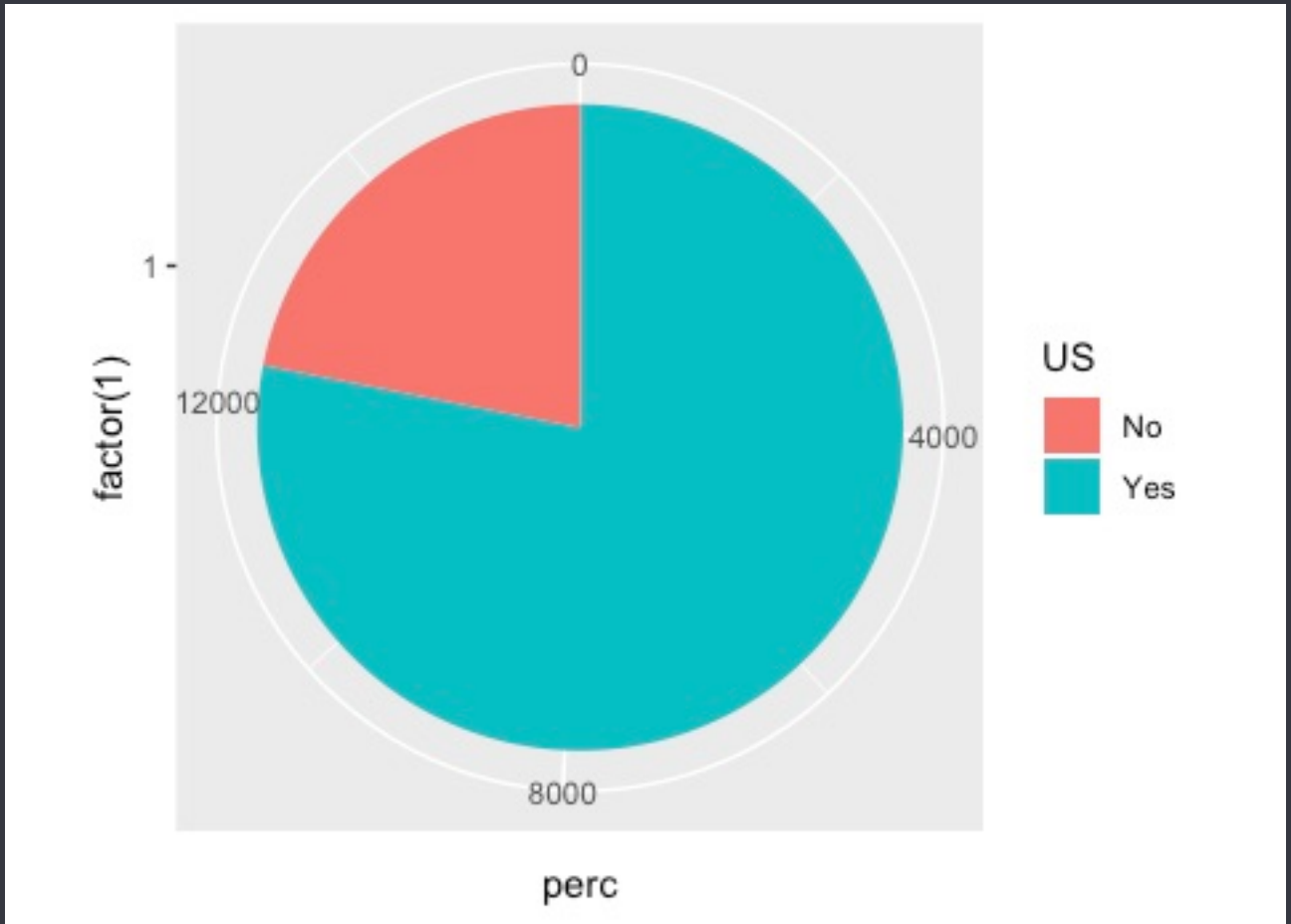
本資料共演團體機制表現



本資料策展人是否為場地擁有者

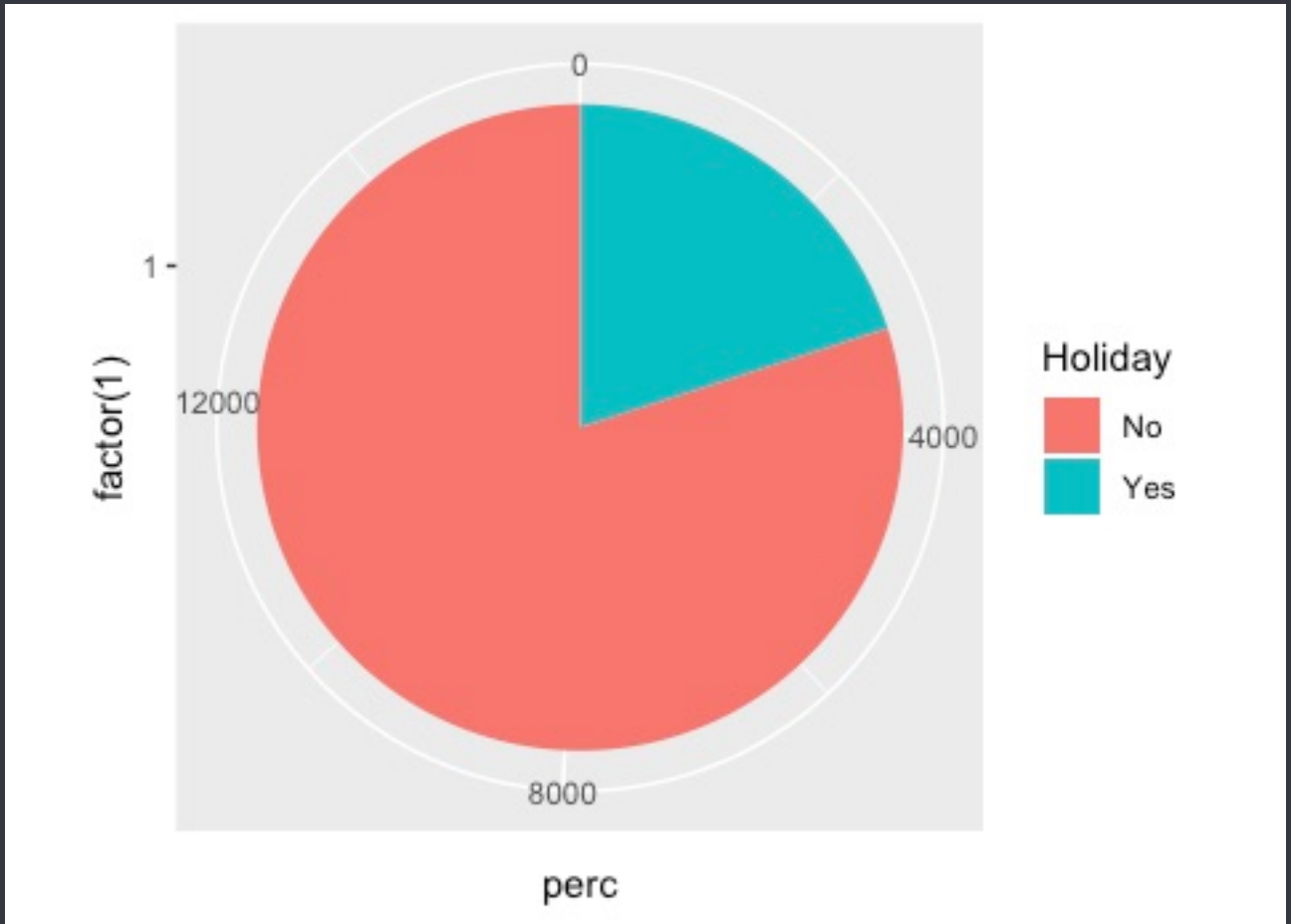


是否辦在美國

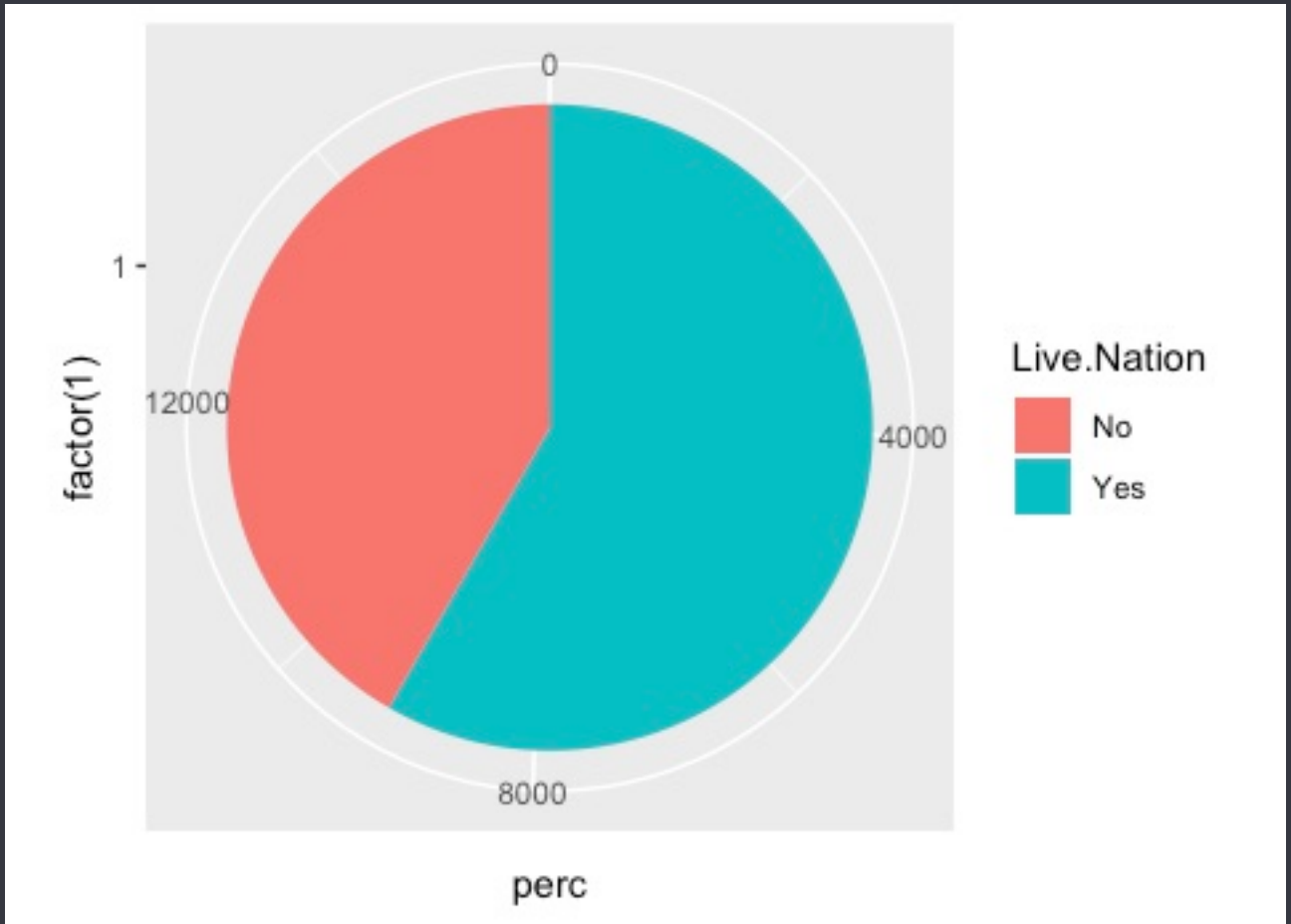


演唱會是否為假期

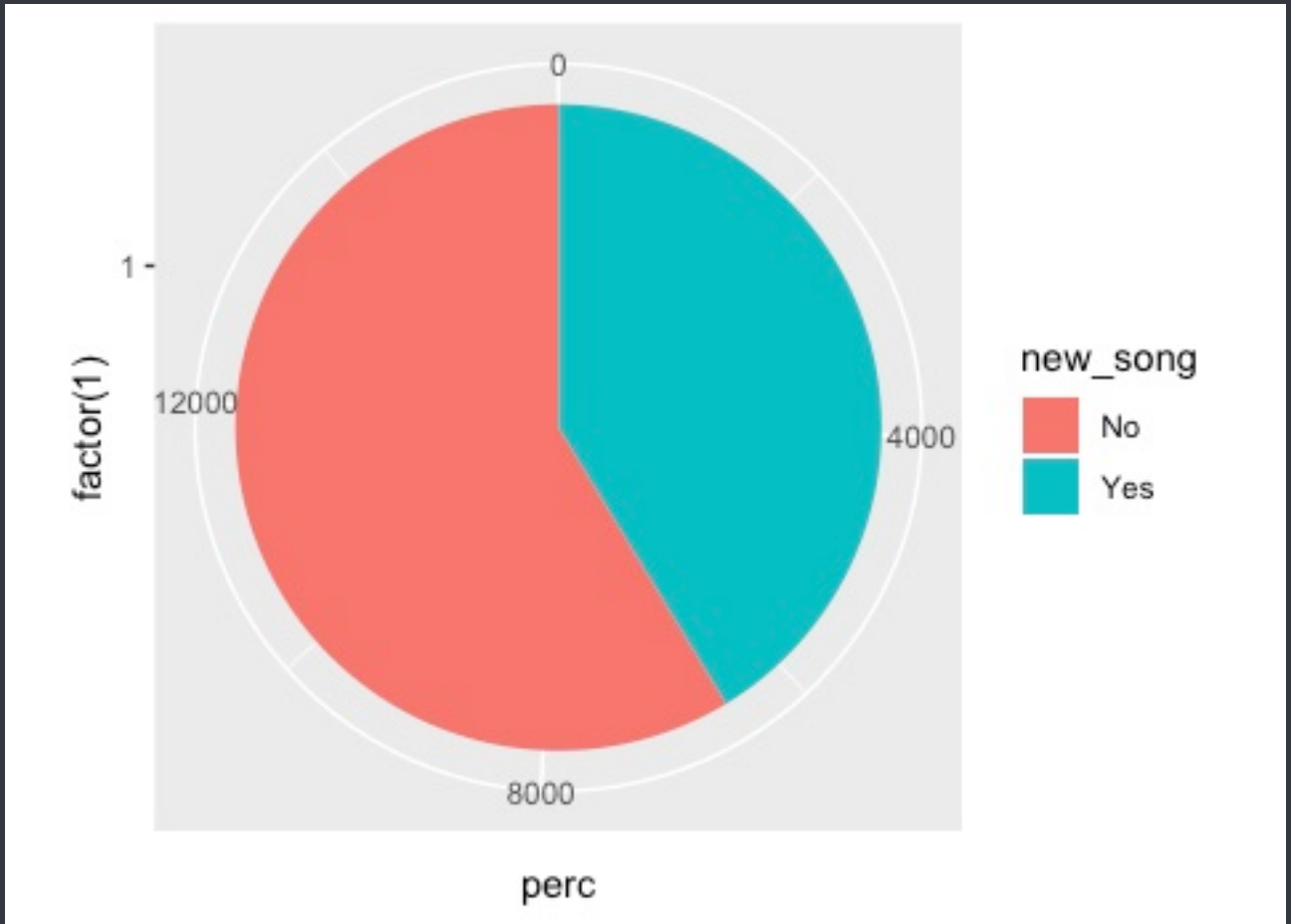
#問題：週末並未計入為假期，且所有連假加總日數本來就比週休二日的總數少！



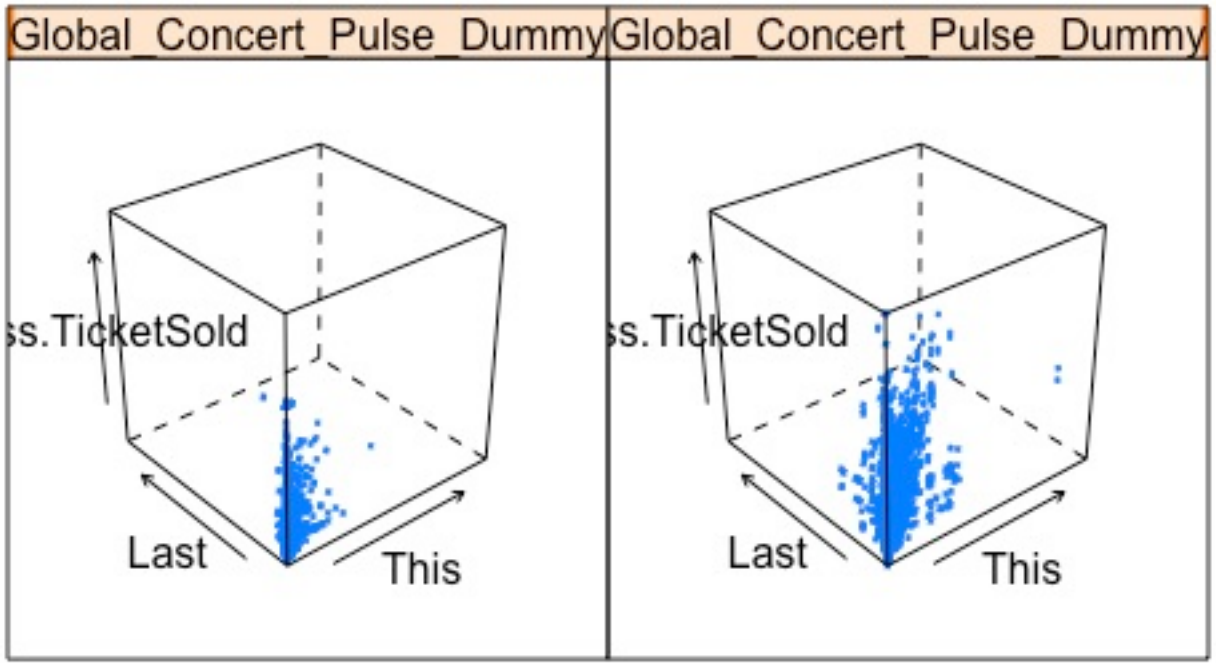
策展人是否為Live Nation



演唱會前是否發表新歌



#\是否登過百大排行，來區分其上週點擊率與本週點擊率觀察對每票收益的影響



圖片：Oasis

什麼是機器學習？

機器學習大方向

圖形辨識、群集分析

- 類神經網路
- 決策術
- 感知器
- **支持向量機 (SVM)**
- AdaBoost
- 貝式分類器

迴歸分析、統計分析

- 線性迴歸
- 高斯過程迴歸
- 最近鄰居法 K-NN

機率密度

- 最大期望演算法
- 機率圖模型
- **套索算法 （Lasso Regression）**

支持向量機 SVM

動物辨識器

給機器（模型）一大堆動物的圖片，告訴它每種動物的特徵是怎麼樣的。

給它一張圖來做測試，問機器這是什麼動物？

貓：5

狗：3.2

豬：1

也有犯錯的時候

它根本沒看過這種動物，或者特徵不夠明確，容易混淆。

這到底是馬芬蛋糕還是吉娃娃！？

圖片：Rejjie Snow

Lasso Regression (套索算法) 與 SVM (支持向量機)

Lasso Regression 套索算法

最小絕對值收斂

特徵選擇、正則化，擬合最適模型，做出最佳子集選擇。

- 最初應用於統計上的最小二乘法。
- 增強統計模型的預測精確度與可解釋度。
- 強制讓迴歸係數絕對值之和小於某固定值。
- 強制讓垃圾變數的迴歸係數為0。

- 變數欄位從48欄刪減至34欄。

$$\sum_{j=1}^{p} |\beta_j| \leq t \qquad \min_{\beta_0,\beta} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\}$$

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 \right\} \text{ subject to } \|\beta\|_1 \leq t.$$

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

https://github.com/ritalinyutzu/Thesis

刪減後資料變數

包含34個變數

- Shows：該演唱會辦了幾場表演。
- Year：演唱會舉辦的年份。
- Month：演唱會舉辦的月份。
- Promoters：演唱會之策展人。
- TicketSold：該場演唱會賣出幾張票。
- Capacity：該場地能容納之人數。
- Percentage：售票率。
- SoldOut：是否完售，為虛擬變數。
- US_Gross：演唱會票房總收益。
- min_price：該場演唱會之最低票價。
- US_Gross.TicketSold：票房總收益除以總售出票數，為每票收益。
- Supporting：是否有共演團體（暖場表演）。
- Global_Concert：是否有登上百大演唱會排行榜，為虛擬變數。
- Global_Concert_Times：登榜次數。
- InHouse_Promotion：場地擁有者是否同時為策展人。
- no_promoter：策展人的數量。
- US：是否舉辦在美國，為虛擬變數。
- Canada：是否舉辦在加拿大，為虛擬變數。
- Europe：是否舉辦在歐洲，為虛擬變數。
- UK：是否舉辦在英國，為虛擬變數。
- SouthUS：是否舉辦在美國南部，為虛擬變數。
- Oceania：是否舉辦在大洋洲，為虛擬變數。
- Date_Whichday：舉辦演唱會的日期為哪一天。
- Live.Nation：策展人是否為Live Nation，為虛擬變數。
- Holiday：舉辦演唱會該天是否為假期，不包含週末，為虛擬變數。
- Open.x：Alphabet 當天開盤價格。
- High.x： Alphabet 當天漲停價格。
- Volume.x：Alphabet 當天成交量。
- Low.y：Live Nation 當天跌停價格。
- Volume.y：Live Nation 當天成交量。
- Last：舉辦演唱會前，Youtube歌曲瀏覽量。
- This：舉辦演唱會後，Youtube歌曲瀏覽量。
- Change：舉辦演唱會前後，Youtube歌曲瀏覽量之變化量。
- new_song：該場演唱會前是否發布新歌。

SVM 支持向量機

將所有變數進行特徵化

用統計風險最小化原則來估計分類的超平面，找到一個決策邊界，使類別間的邊界最大化。

把所有變數進行特徵化，再進行SVM 來預測。

- 將每票盈利(US_Gross.TicketSold) 分為五類。
（依照Q3、Mean、Median、Q1、Other）

“5” ->Q3
“4” ->Mean
“3” ->Median
“2” ->Q1
“1” ->比Q1還低

- 將所有虛擬變數及integer 欄位都轉換為Factor。
- 其他Numeric 變數依照上述方法也進行分類。

結果：預測精準度 66.11337%

https://github.com/ritalinyutzu/Thesis

圖片：Avenged Sevenfold

SVR Regression

不需要將所有變數進行特徵化，
直接以數據來進行SVR 來預測。

SVR Regression

不需將變數特徵化

用統計風險最小化原則來估計分類的超平面，找到一個決策邊界，使類別間的邊界最大化。

https://github.com/ritalinyutzu/Thesis

圖片：Lamb of God

2SLS 兩階段最小平方法

2SLS 模型

y1 = b0 + b1*y2 + b2*y2 + . . .b24*y25 + b26*z26+ . . . + b33*z33 + u

y2 = π0 + π1z1 + π2z1 + . . . π33z33 + v

Residual standard error: 49.75 on 15736 degrees of freedom
Multiple R-Squared: -0.9463
Adjusted R-squared: -0.9493

Wald test: 112.7 on 24 and 15736 DF
p-value: < 2.2e-16
F檢定非常顯著

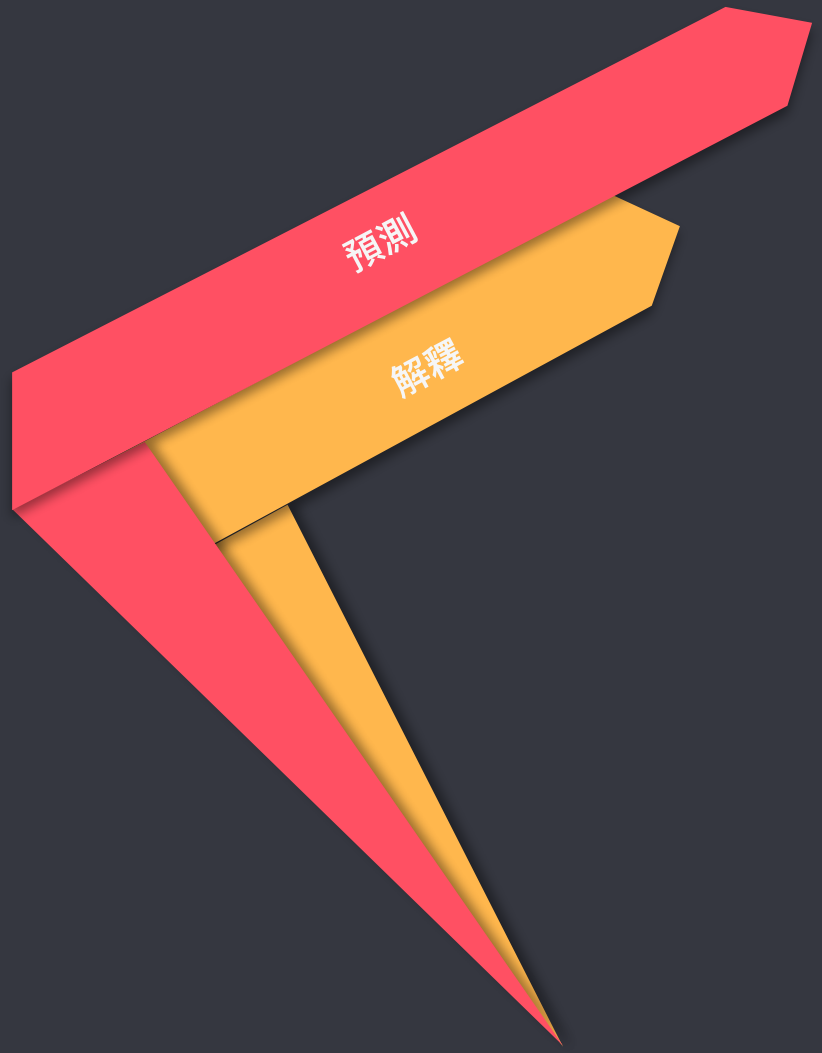
Youtube 網路聲量無法解釋演唱會票房之每票收益。

然而 new_song 有非常正向顯著，
顯示出觀眾在喜愛的音樂人「發佈新歌」時傾向去看演唱會。

https://github.com/ritalinyutzu/Thesis

圖片：Maroon 5

結論



解釋與預測之間的抉擇

若希望完全解釋，則失去預測性；若著重於預測，則會忽略對於資料的解釋。

- SVM Accuracy：預測精準度 66.11337%
- SVR Regression RMSE： 10.32287
- 2SLS F-test： 2.2e-16

- 註：RMSE為預測誤差，是MSE開根號，可理解為殘差。

- **舉辦演唱會前、後之 Youtube 點擊率 ，能夠預測未來的每票收益。**
- **舉辦演唱會前、後之 Youtube 點擊率，無法完全解釋每票收益。**

- **結論：Youtube 網路聲量並非對於演唱會票房市場的最直接影響，僅能夠作為一參考依據。**

後續進行內容

加入2019年2、3月資料

目前資料為2017/02-2019/01

反覆確認內容是否有誤

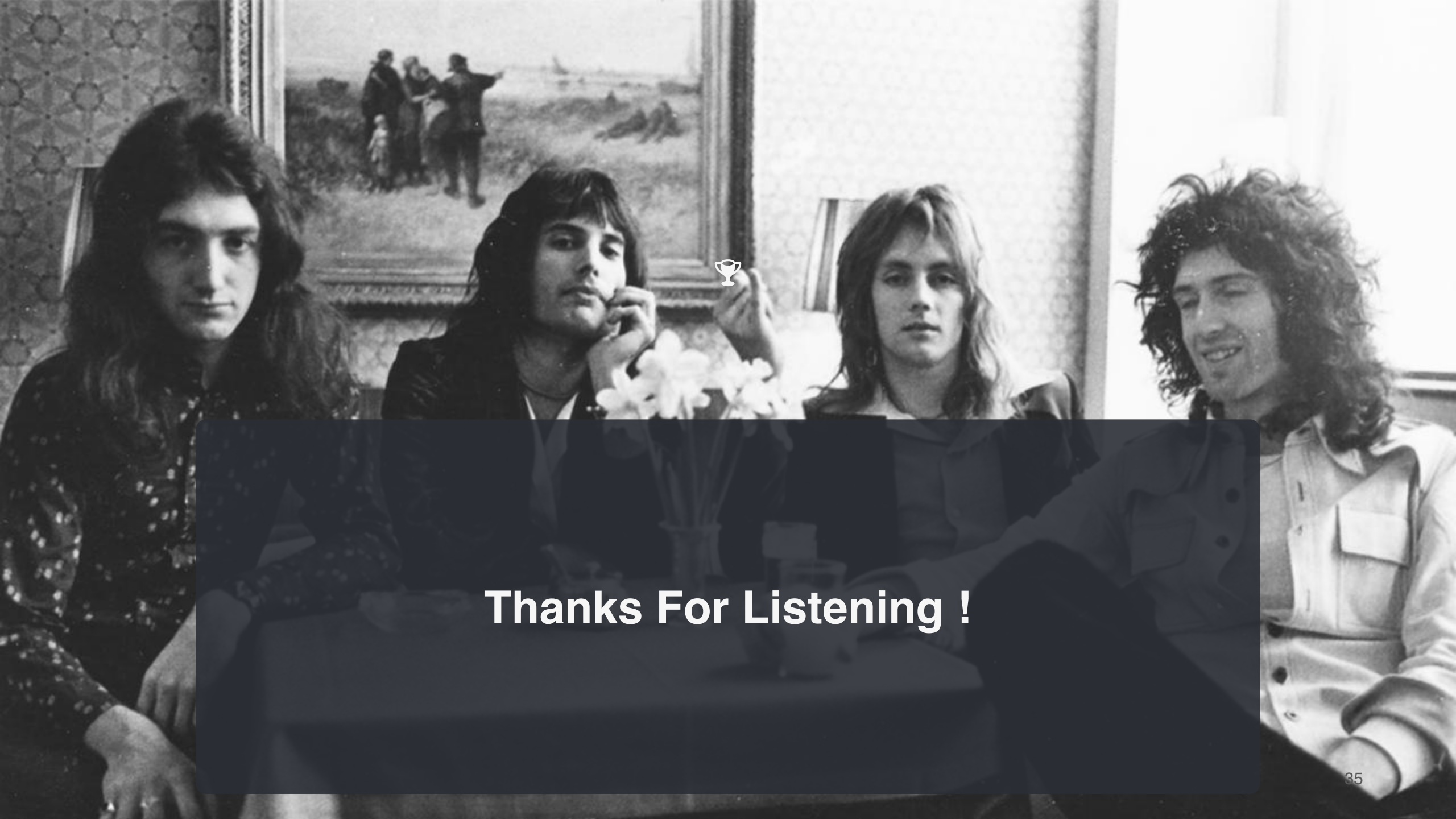
有問題應快速Debug，沒問題就趕快畢業好好找工作！！！

提升SVM模型精準度

將模型預測調整至70%

書寫論文內容

書寫論文正式內容，並於五月參加論文研討會。



Thanks For Listening !