KKBOX Churn 分析文件

資料集:

(1) Members.csv :
　性別、年齡、居住城市、註冊方式、開始日期、到期日
(2) Transactions.csv :
　交易方式、交易金額、交易日期、是否自動更新、是否取消續約
(3) Userlogs.csv :
　聽歌總秒數、聽歌歌曲數、聽歌長度(25,50,75,985,100)
(4) Train.csv : 該帳號是否流失

Userlog.csv 目前處理方式:

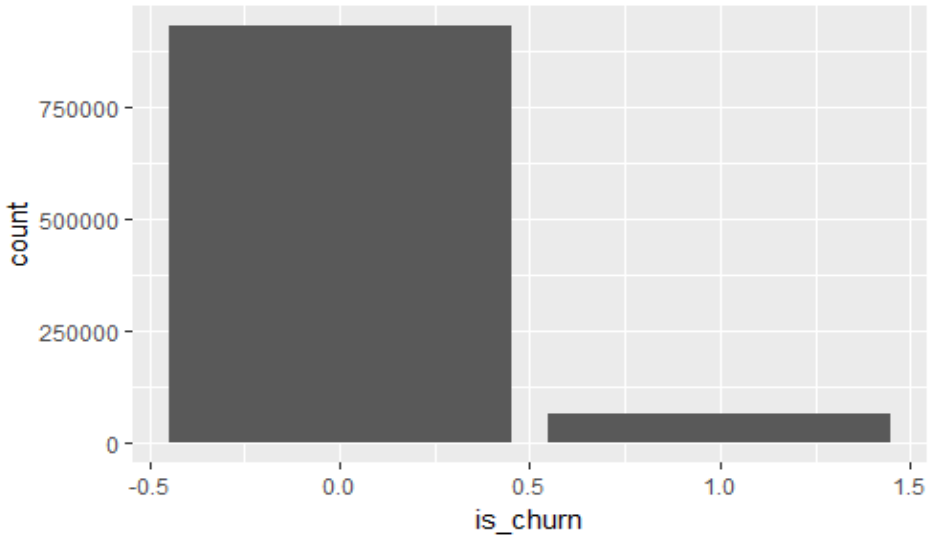
將四億筆資料的 userlog.csv 分成五十份，每一份 distinct 過後，

再重新 bind 起來，最後比數約 95 萬筆。

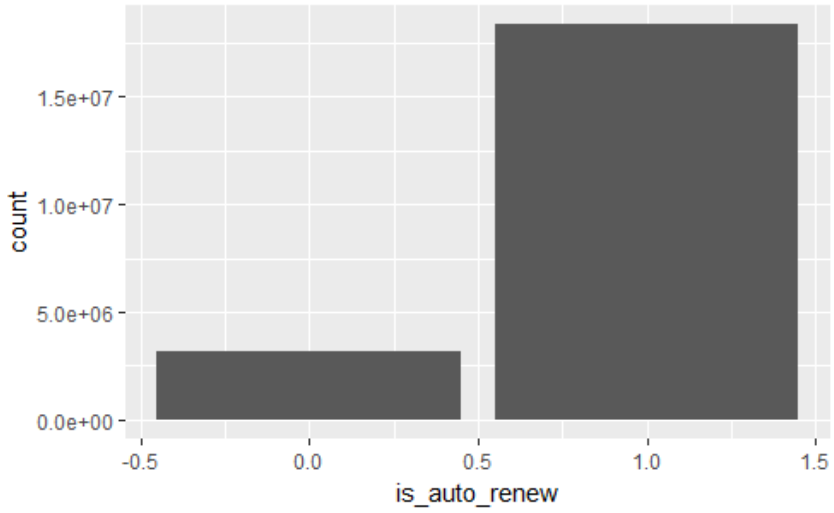
Distinct 缺點：可能會流失使用者習慣的改變之資料

探索性分析:

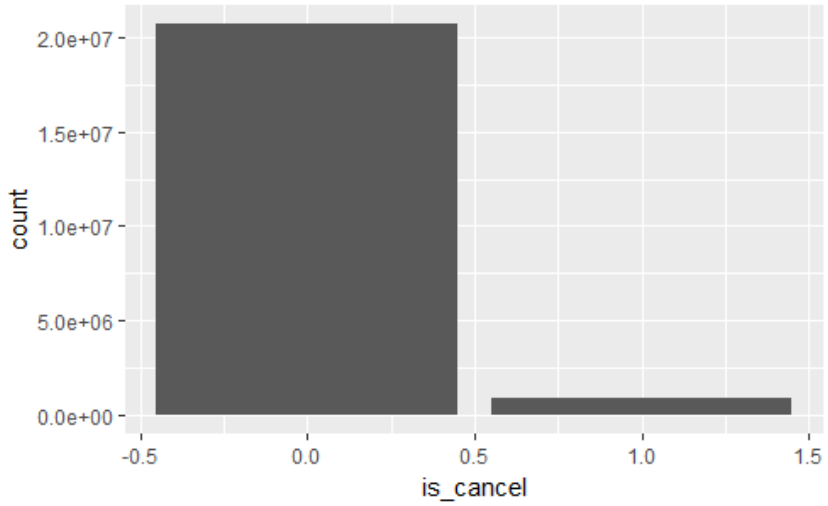
(1) Churn 為不平衡資料 (0.94 0.06)



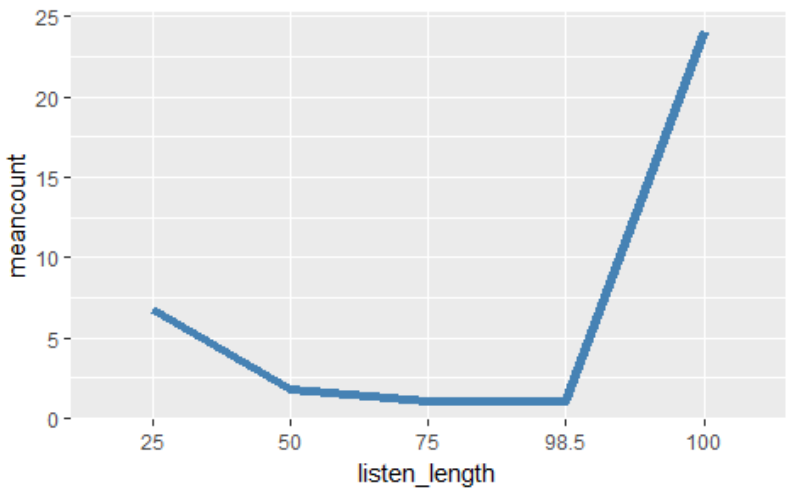
(2) 大多數人會自動更新 #0.148 vs 0.852



(3) 少數人會結束合約 #0.96 vs 0.04



(4) 聽歌長度的習慣 不是聽一下 就是聽完



目前準備分析方式:

(1) 產生分析大表:

以 members.csv 為主表，去跟其他 table 做 join 如下:

Console = left_join(members,userlogs)

Console = left_join(Console,train)

建立每個會員的 attribute 後，去跟交易資料作合併，這裡將 transaction 當作左表，因為這樣才能知道每一筆交易資料

Console = left_join(transactions,Console)

(2) 做市場區隔

利用不同年齡層、不同性別、不同交易方式先做初步市場區隔，並利用交易時間間隔建立 RFM。

(3) 找出現況之黃金族群以及熱門商品

利用現況之資料做探索性分析找出對於商品最有黏著度之族群

(4) 會員流失預測分析

利用資料內給的 is_churn 做流失預測，並找出流失最多的族群(可能是年齡層、性別、或是某種消費習慣的消費者)