線上零售業資料分析

演講者: 張順益

大綱

01 資料視覺化

02 自然語言處理

03 模型介紹

04 結論

01

資料視覺化

資料介紹

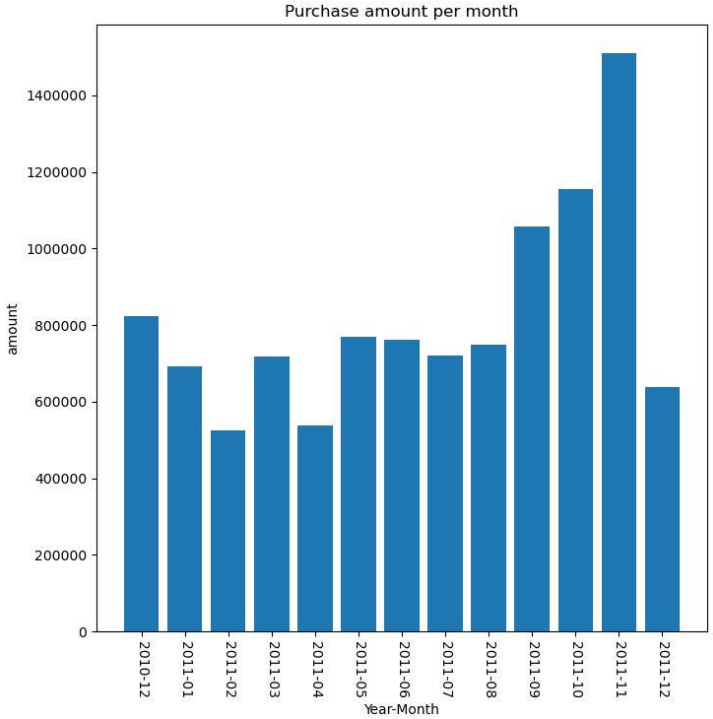
- 發票ID
- 商品ID
- 商品敘述
- 數量
- 購買日期
- 商品單價
- 國家

資料介紹

- 發票ID A開頭代表公司調整呆帳
- 發票ID C開頭代表折扣所以數量可能為負
- 其餘數量為負可能為退貨

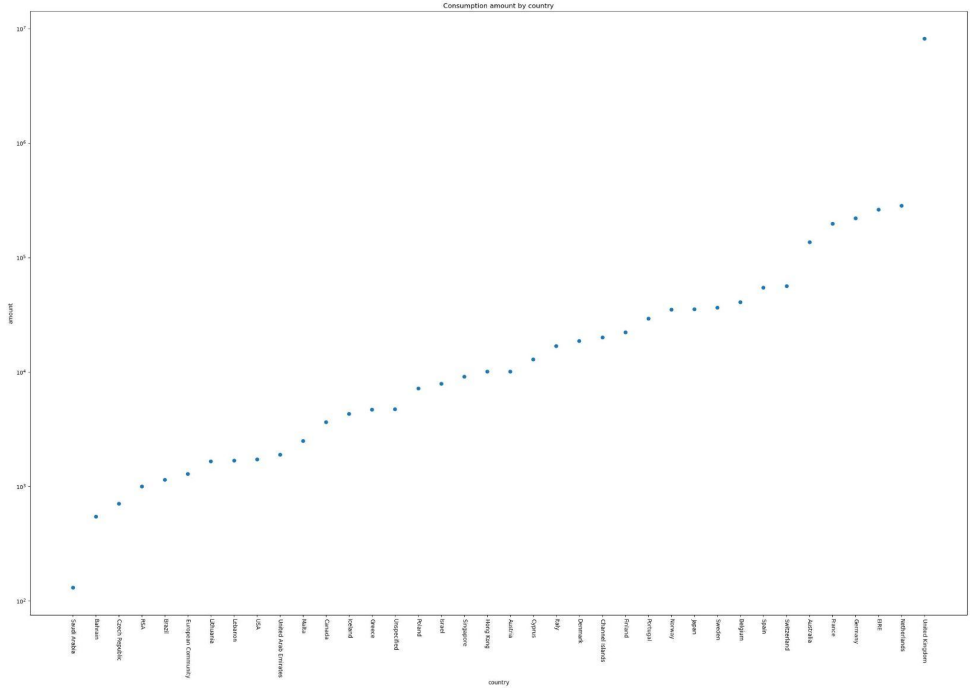
資料視覺化

- 可以看出年底銷售額較好
- 2011-12資料只到9日所以銷售額較低
- 只有一年數據



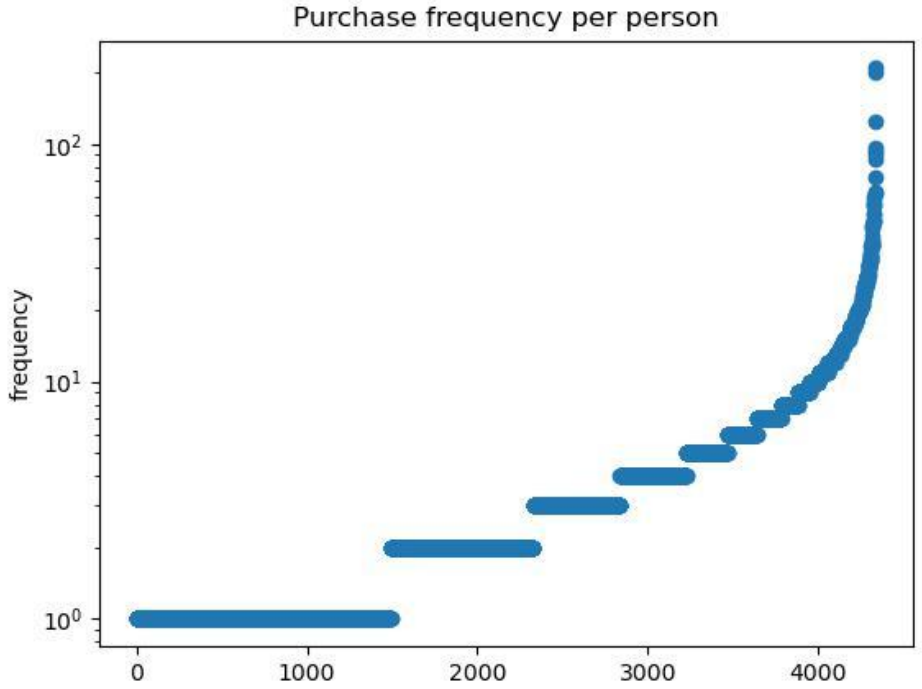
資料視覺化

- 沙烏地阿拉伯為銷售額最低的國家，英國為最高的國家
- 且從此圖可知歐洲國家銷售額都較高中東地區則較低



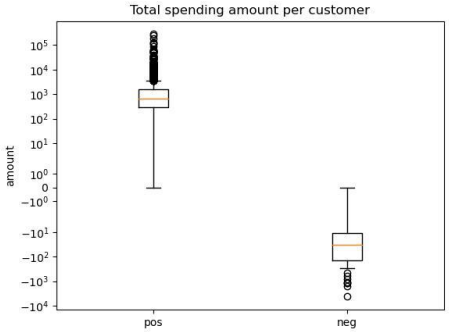
資料視覺化

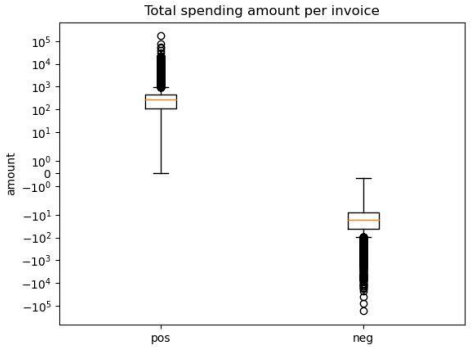
- 可以看出大部分消費者一年消費次數並不高
- 僅有少數過百次



資料視覺化

- 上圖可以得知平均每人一年銷售額為1000左右且大部分集中在附近而最高則有過100000
- 而負的部分則大約集中於-100左右





- 下圖可知大部分單筆消費數百最大值則有過100000
- 而負的部分則集中於數十左右

02

自然語言處理

前處理

- 詞性還原
- 去除標點符號及常見停用詞
- 進行斷詞

自然語言

- 上圖為英國前五大詞出現次數，

```
[('set', 55439),
 ('bag', 47031),
 ('red', 37691),
 ('heart', 37075),
 ('vintage', 30617)]
```

下圖為沙烏地阿拉伯前五大詞

出現次數。

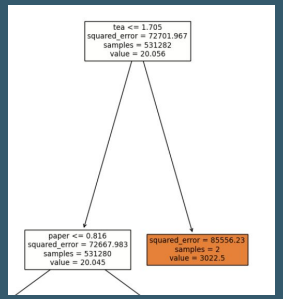
由圖可以推斷出英國主要出現形容詞等看似較

精緻的物品反觀沙烏地阿伯則主要為日常用

```
[('glass', 3), ('jar', 3), ('plasters', 3), ('tin', 3), ('assorted', 1)]
```

自然語言

- 建立tf-idf並刪除缺失值過高的詞
- 經由tf-idf矩陣建立回歸樹尋找影響銷售額的詞彙
- 例如由圖可知tea的tf-idf 較高
- 的情況下銷售額會較高



03

模型介紹

RFM 模型介紹

- 透過最近購買時間，購買次數，購買金額，將顧客進行分群
- 按照上述3者各自分為三個部分，最近購買時間越近評分越高，購買次數越多評分越高，購買金額越高評分越高
- 共分成3 * 3 * 3 = 27群

RFM 模型介紹

- 評分(3, 3, 3)的比例約為16.5%而(1, 1, 1)的比例約為17.1%，可知高價值客戶及需挽留客戶佔資料中最大比例
- (2, 1, 1)比例為8.3%表示一段時間前有來消費，但還處於認識品牌的階段，沒持續消費，所以可以繼續追蹤
- (2, 3, 3)比例為7.6%左右為消費金額跟次數都很高但近期沒消費故可以尋找誘因讓他們回歸消費(1, 1, 2)比例為7.1%左右為有一定消費能力但品牌黏著度不高因此透過活動增加黏著度，讓他們持續消費

04

結論

結論

- 目前客戶重心在歐洲，因此在深根歐洲市場的同時可以開發其他地方
- 許多購買次數較低的客戶表示網站對新進顧客吸引力不足，或許可以透過活動、折價券等等增加客戶持續消費的動力
- 透過自然語言分析找出當tea tf-idf較高能提供較高銷售額，paper tf-idf較低銷售額也會較高，可能表示茶類的銷量較好而紙由於售價較低因此營業額較低

未來展望

- 可以收集更長時間的資料以了解整體淡旺季
- 可以協助蒐集用戶資料以及更多商品相關資訊以預測改善銷售額以及增進客戶黏著度等等
- 自然語言處理後的預測依然不夠準確原因由於方法並不夠好，且敘述資料往往是對商品的陳述可能與銷售額關聯不夠強烈

THANKS

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, infographics & images by Freepik and illustrations by Stories