資料分析與基石

Store Sales - Time Series Forecasting

Use machine learning to predict grocery sales

第 11 組
何子安 E44065020｜蔡東霖 F74071166｜林千祺 N46084036｜吳定洋 F74076213

Outline

- 任務簡介
- 資料概述
- 資料分析
  - oil.csv
  - transaction.csv
  - stores.csv
  - holiday.csv
- 總結
- 未來工作
- 參考資料

任務簡介

- 預測雜貨業的銷售額
- 輸入資料:
  - holiday_events.csv
  - oil.csv
  - stores.csv
  - transactions.csv
  - train.csv
- 輸出: 該筆項目的銷售額(連續值)
- 評估標準
  - Root Mean Squared Logarithmic Error (RMSLE)

資料概述

- Train.csv（資料筆數 3000888 筆，沒有缺失值）

Column	Range/Unique	Description
id	3000888	索引值
date	2013.01.01~2017.08.15	日期
store_nbr	54 間	商店編號
family	33 種產品	產品類別
sales	124717	該產品於該商店當日的總銷售額
onpromotion	741	該產品於該商店當日促銷的數量
備註	發薪日為每月15日及月底；2016.04.16 厄瓜多大地震，地震發生後數週，人們齊心協力捐贈水和其他急需物資，極大地影響了超市的銷售。	

資料分析

- **Transactions.csv**
- Oil.csv
- Stores.csv
- Holiday_events.csv

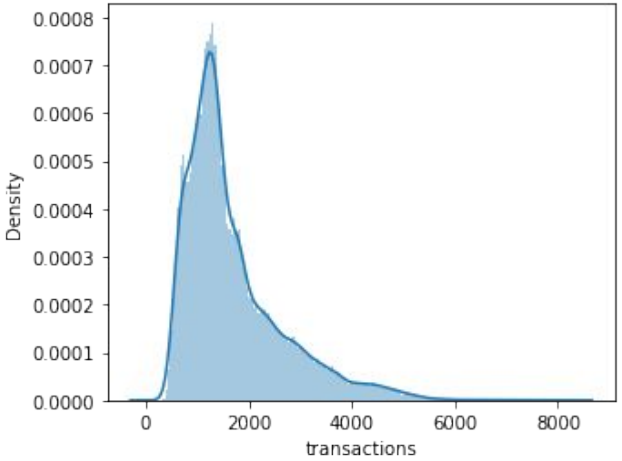
Transaction資料分析

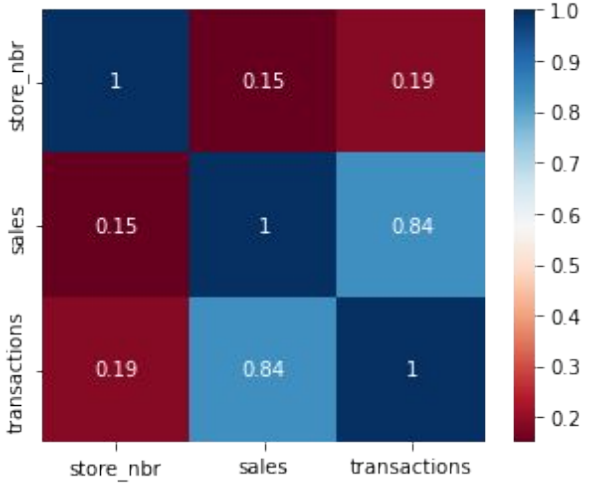
- transaction.csv（資料筆數 90936，沒有缺失值）
- 將 date 資料類型改為 datetime
- 將 train 資料根據日期，將各店的銷售量進行加總，產生每日各店銷售量
- 將 transaction.csv 與 train.csv 進行合併，得到新的 column
- 合併後 transaction 欄位會產生 7448 個缺失值

Column	Range/Unique	Description
date	2013.01.01 ~ 2019.08.15	日期
store_nbr	54 間	商店編號
transaction	5 ~ 8359	交易量

Transaction資料分析

- 將 transaction 與 sales 進行相關性分析
- 相關係數 ＝0.8374
- 資料型態為單峰、右偏，表示偏低的資料較多





Transaction資料分析

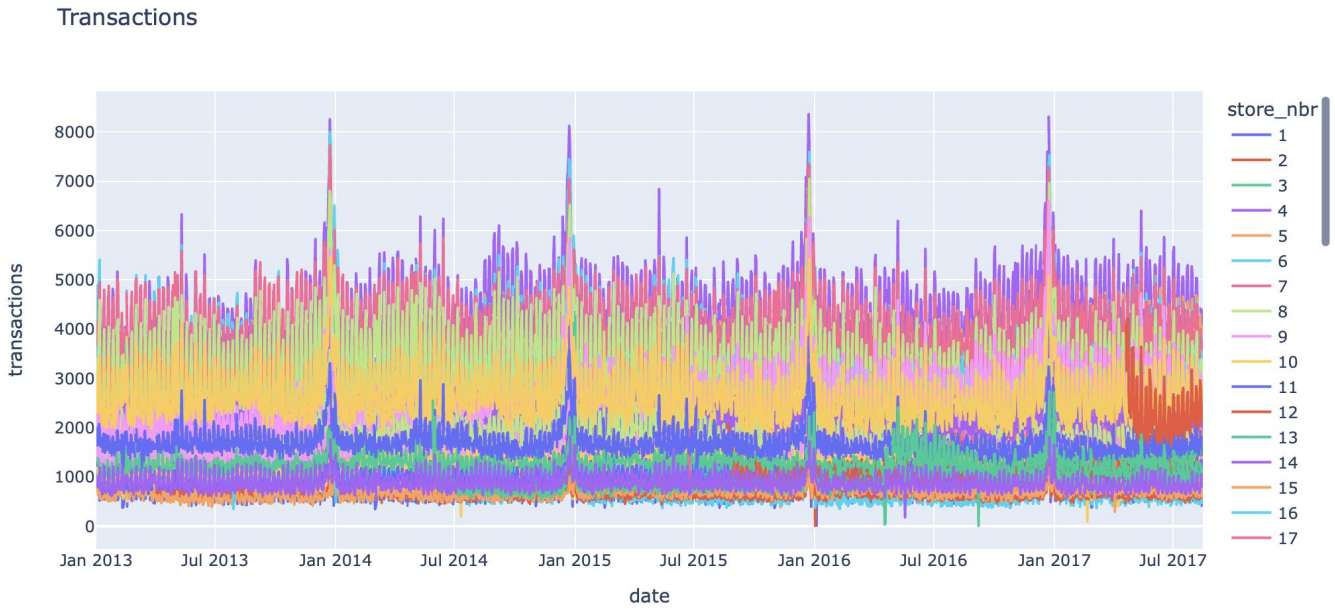
- 繪製散佈圖
- 可以看到離群值大多分布在偏低的 transaction 資料中
- 離群值偏高

Transaction資料分析

- 繪製箱型圖
- 12月有最多的離群值，且資料較離散
- 12月的交易量最大

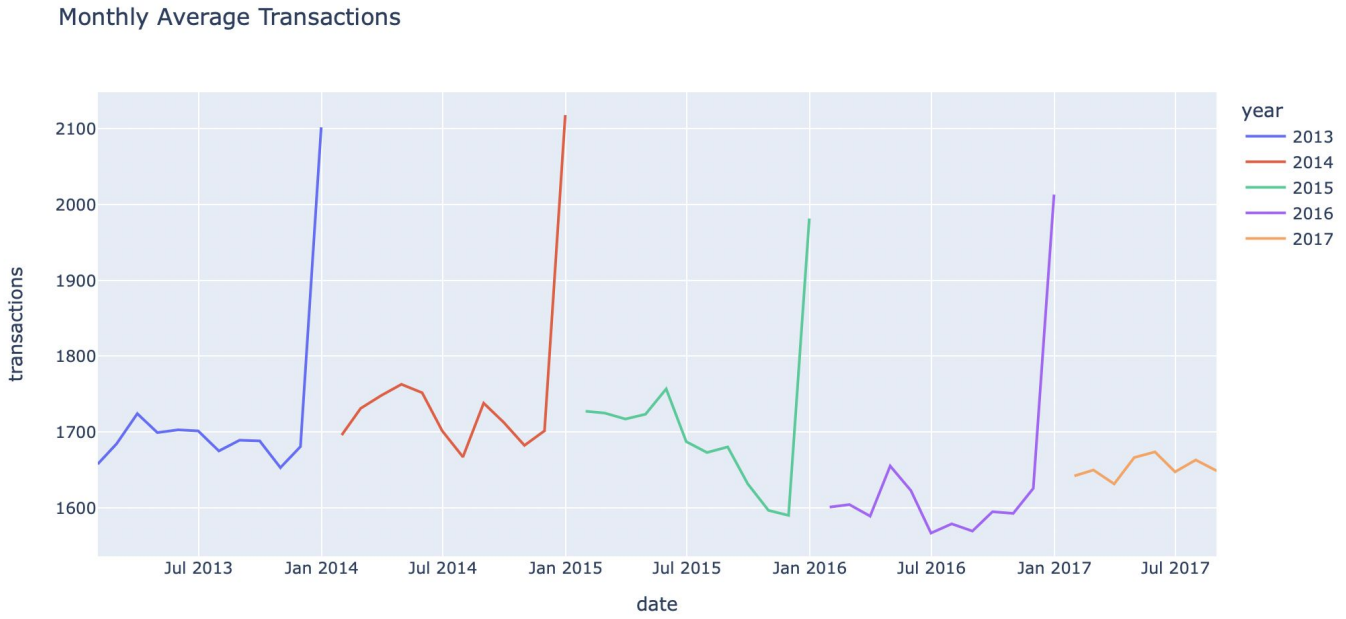
Transaction資料分析

- 繪製各商店交易量隨時間變化的折線圖
- 可以發現每家店在12月時，交易量皆會顯著上升



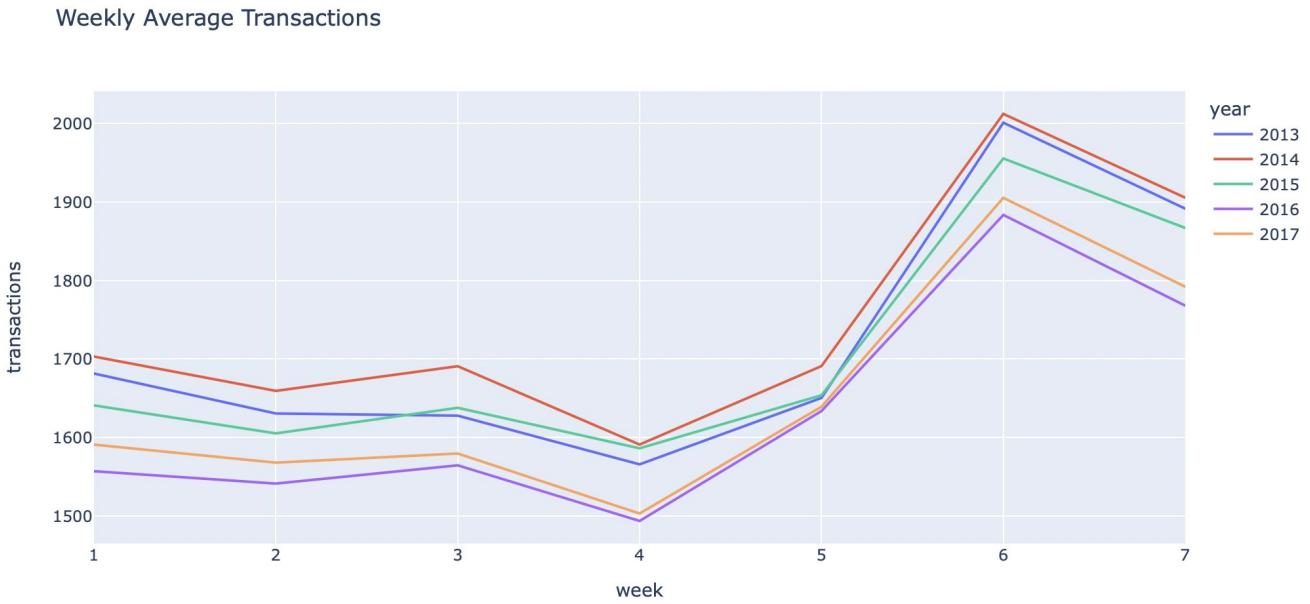
Transaction資料分析

- 繪製每年的月交易量變化折線圖
- 可以看出每月的交易變化趨勢是很相似的，並且隨著年份緩慢遞降
- 2016 年的交易量最低，在4月份時有一個小高峰，推測與地震有關



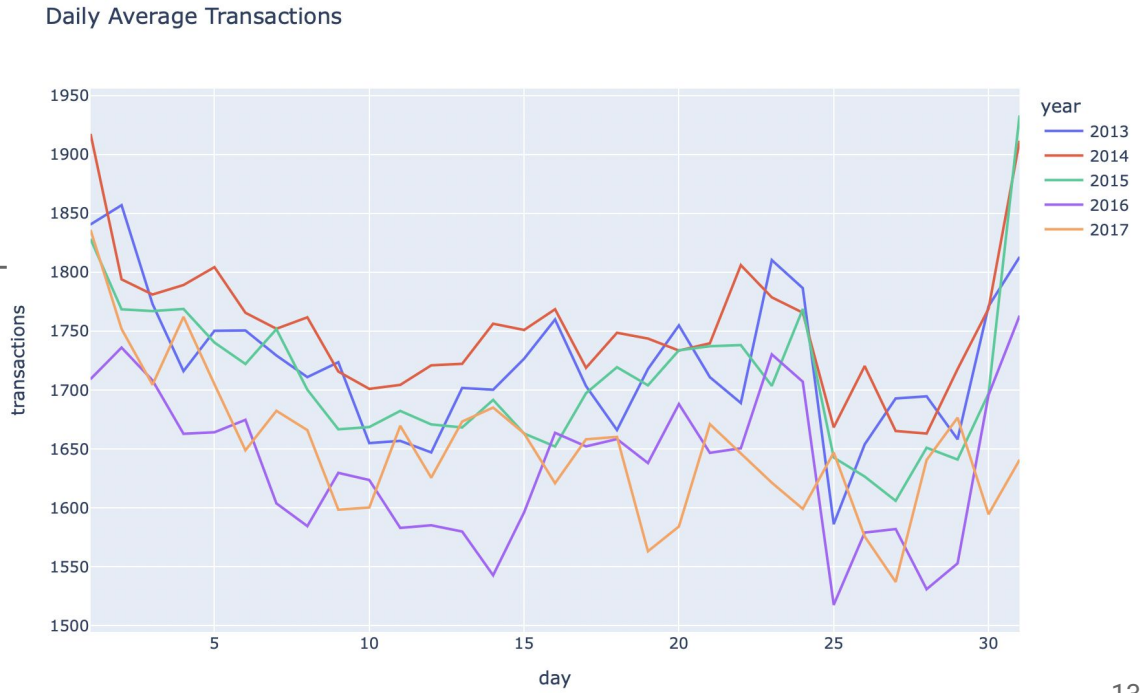
Transaction資料分析

- 繪製每週交易變化折線圖
- 可以看到在週末時，交易量上升，週六最高
- 週四交易量最低



Transaction資料分析

- 繪製單月中的每日平均交易量變化圖
- 可以看到月初與月末交易量最高
- 緩慢遞減至15日左右，會再小幅度上升
- 每月25日左右是交易量最低的時期
- 推測與發薪的日期有關

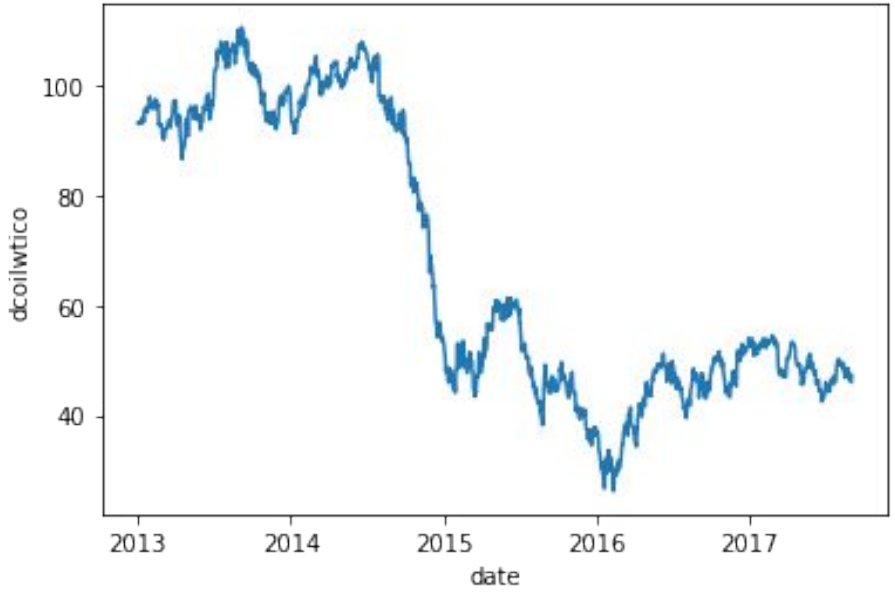


資料分析

- Transactions.csv
- **Oil.csv**
- Stores.csv
- Holiday_events.csv

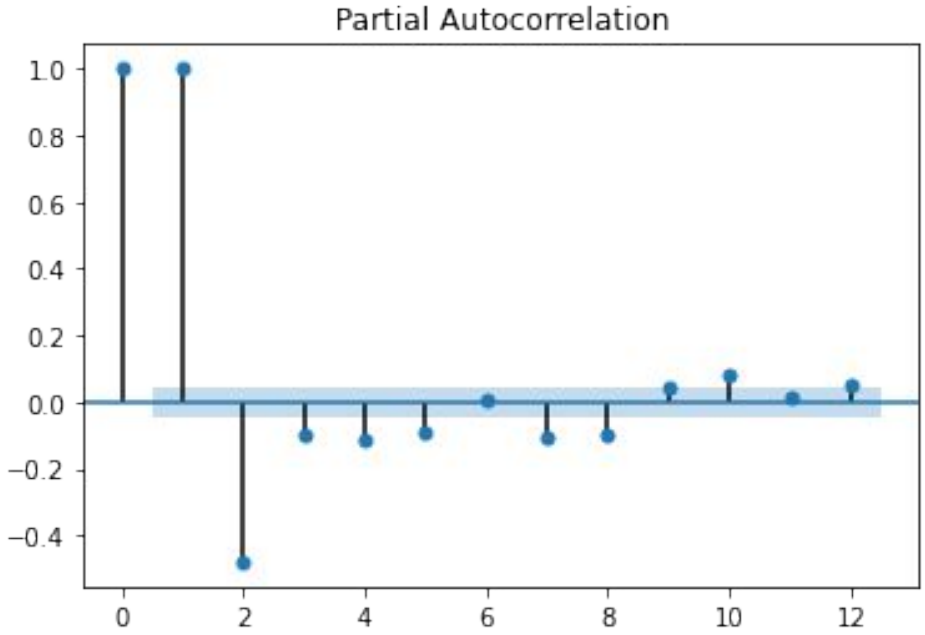
oil資料分析

- 油價與時間的關係圖
- 油價長期來看是漸低的
- 2016是低谷
- 2013-2014間是高峰



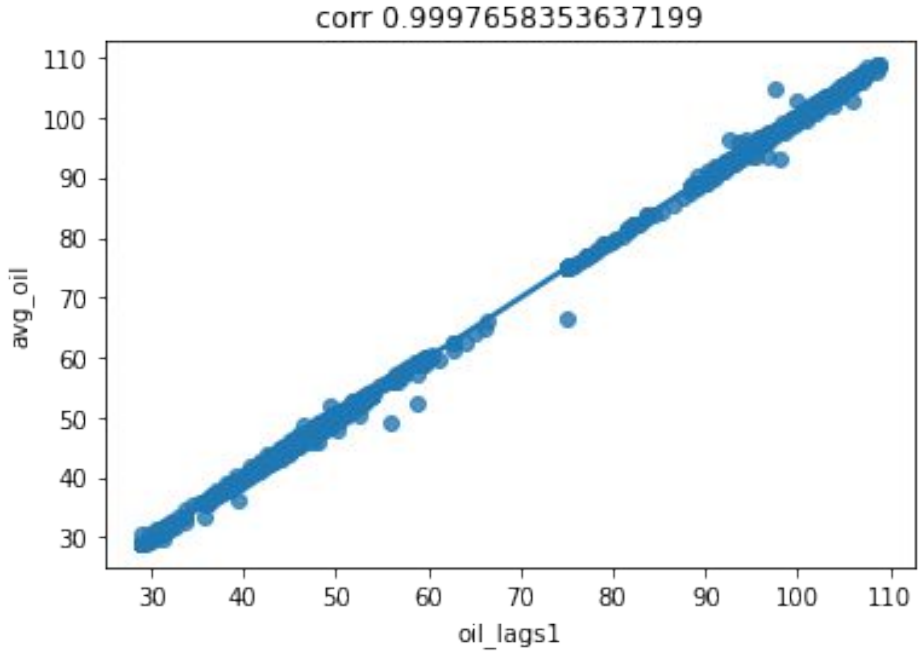
oil資料分析

- Oil的自相關圖-滯後圖
- 產生滯後的最大值高達 5
- 油價漸趨平緩



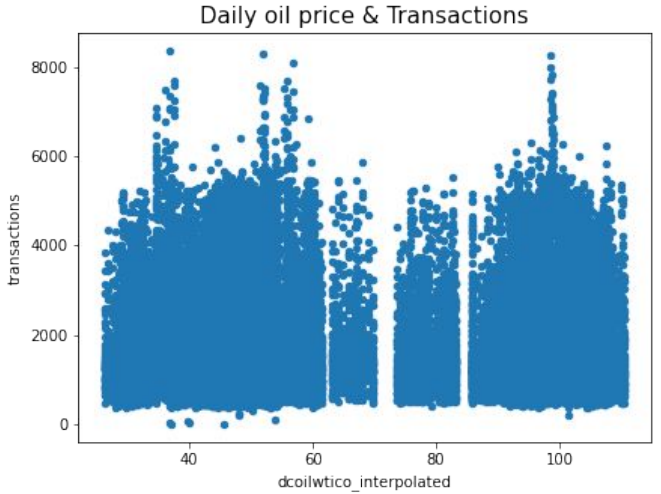
oil資料分析

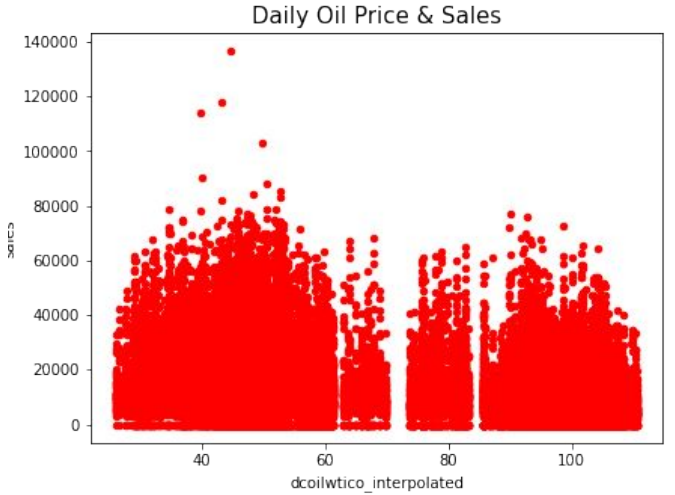
- 滯後平均油價vs及時油價的關係
- 滯後油價可以預測未來的油價



oil資料分析

- 油價vs交易量
- 油價vs銷售量
- Correlation: transactions 0.04
- Correlation: sales -0.30
- 可以看出兩者沒極大相關





資料分析

- Transactions.csv
- Oil.csv
- **Stores.csv**
- Holiday_events.csv

stores資料分析

- stores.csv（資料筆數 54 筆，沒有缺失值）
- 關於這些店的 metadata

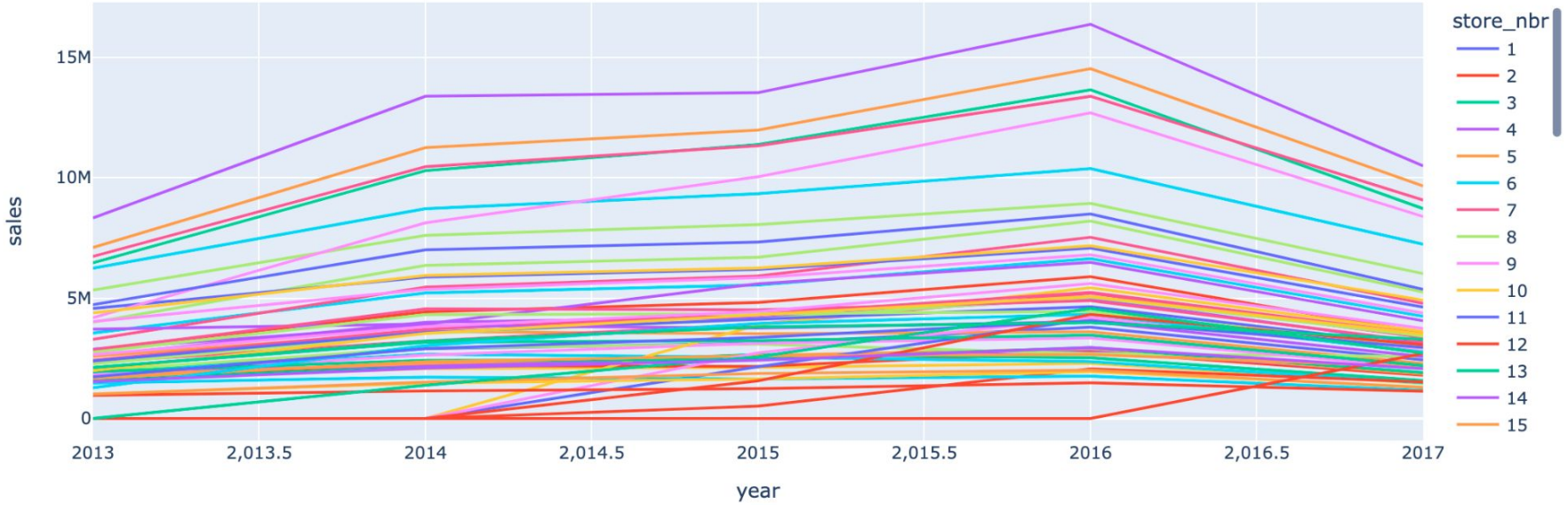
Column	Unique	Description
store_nbr	54 間	資料中涵蓋了 54 間店的其他資料
city	22 個城市	這些店分別在 22 個城市內
state	16 州	這些店分佈在 16 個州中
type	5 類	共 5 種不同類型的商店
cluster	17 種	可分為 17 群相似的商店

stores資料分析

- 從該相關矩陣可以發現sales 與 onpromotion 有比較高的相關性
- 且在不同的店面都有相似的情況

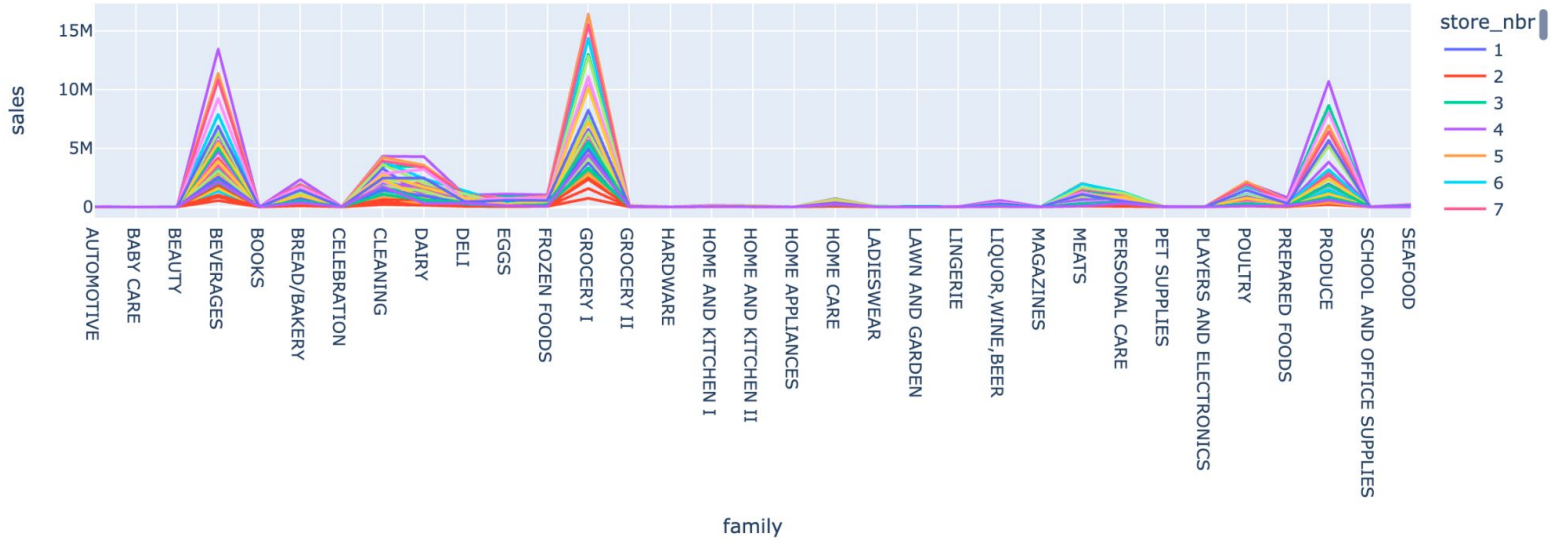
stores資料分析

- 幾乎所有的店面在 2013 - 2015 年中 sales 都有趨增的情況
- 且都在 2016 年的 sales 有下降的趨勢



stores資料分析

- 所有店家的熱賣商品都集中在這一些品項
- GROCERY 1, BEVERAGES, PRODUCE, CLEANING, BREAD/BAKERY



資料分析

- Transactions.csv
- Oil.csv
- Stores.csv
- **Holiday_events.csv**

holidays_events.csv 初步觀察

holidays_events 這份資料有 6 個 column：

1. date：日期。
2. type：只有[‘Holiday’ ‘Transfer’ ‘Additional’ ‘Bridge’ ‘Work Day’ ‘Event‘]這6種。
3. locale：只有[‘Local’ ‘Regional’ ‘National’]這3種。
4. locale_name：有兩種可能，一個是州的名字，一個是城市名字。
5. description：對這個 holiday 或 event 的描述，判斷是什麼 Holiday 的依據。
6. transferred：true or false，這天holiday是否有被改移到別天，
    a. 如果 = true，代表原本這天的假期被移到別天去了，這天等於沒有放假。

holidays_events type column整理

type中有['Holiday' 'Transfer' 'Additional' 'Bridge' 'Work Day' 'Event']這6種，但我們只在意的是在data這個特定日期中，是否有放假或是特別活動(會影響購物的，ex:黑色星期五)。

<Step1>

將type先分成Holiday、Work Day。首先處理處理Transfer項。

```
# drop transferred column and reset index and store the result in tr1(all holiday original)
tr1 = holidays[(holidays.type == "Holiday") & (holidays.transferred == True)].drop("transferred", axis = 1).reset_index(drop = True)
# drop transferred column and reset index and store the result in tr2(all transferred holiday)
tr2 = holidays[(holidays.type == "Transfer")].drop("transferred", axis = 1).reset_index(drop = True)
```

將type==Holiday、transferred == True的放到tr1 (沒放假)

將type==Transfer的放到tr2 (實際放到假)

```
tr = pd.concat([tr1,tr2], axis = 1)
```

接著將tr1、tr2接起來

	date	type	locale	locale_name	description	date	type	locale	locale_name	description
0	2012-10-09	Holiday	National	Ecuador	Independencia de Guayaquil	2012-10-12	Transfer	National	Ecuador	Traslado Independencia de Guayaquil

最後留下正確的放假日期，並且讓type = Holiday即完成，也就是留下綠框的部分。

holidays_events type column整理(Cont.)

接著將type == Aditional 和type == Bridege的資料type都改成Holiday，因為這兩個type都代表有放假。

再將holidays_events分成兩筆資料，work_day、holidays。

```
work_day = holidays[holidays.type == "Work Day"]
holidays = holidays[holidays.type != "Work Day"]
```

<step2>

將holidays分成，**national**、**regional**、**local**、**events**

因為不同地區有不同的holiday，例如，Manta這個city在2012-03-02有地區假期Fundacion de Manta

Cotopaxi這個state在2012-04-01有地區假期Provincializacion de Cotopaxi

	date	type	locale	locale_name	description
0	2012-03-02	Holiday	Local	Manta	Fundacion de Manta
1	2012-04-01	Holiday	Regional	Cotopaxi	Provincializacion de Cotopaxi

而events和national都是國家假期。

結束後目前有**national**、**regional**、**local**、**events**、**work_day**

連結整理完的holidays_events和train data

1.將 train data 依序和national、regional、local、work_day merge，多出4個feature

	id	date	store_nbr	family	sales	onpromotion	city	state	type	cluster	holiday_national	holiday_regional	holiday_local	IsWorkDay
0	0	2013-01-01	1	AUTOMOTIVE	0.0	0	Quito	Pichincha	D	13	Primer dia del ano	NaN	NaN	NaN

2.將所有holiday、event，列出來形成新的feature，例如event共有Black_Friday、Cyber_Monday、Dia_de_la_Madre、Futbol、Terremoto_Manabi這五種，將他們轉為新的features，例如這天剛好是Black_Friday則 == 1，不是則 == 0。

	date	events_Black_Friday	events_Cyber_Monday	events_Dia_de_la_Madre	events_Futbol	events_Terremoto_Manabi
53	2013-05-12	0	0	1	0	0

Holidays也依此方法做，會獲得許多新的feature，Holiday和Event共有50種。

將新獲得的 features 進行AB test

- AB test 中,
  - A group 為**有**此 holiday，發生那天的所以商店所有商品 sales,
  - B group 為**無**此 holiday 發生那天的所以商店所有商品 sales。
- 看有 holiday 和 event 的日子對於 sales 有沒有顯著意義。
- 從中觀察 holiday、event 有無 statistically significant，並在後續優先考慮為訓練模型用的 features。

總結

- **Transactions**
  - 與 sales 之間有顯著相關性，因此將作為模型的輸入資料。
  - 從 boxplot 中可以看到，在每年的十二月份有較多的離群值，
  - 因此最後在做資料篩選時，可以考慮刪除一些離群值，以提高模型正確率。
- **Oil**
  - 似乎跟其他數據沒有極大的關係，但在奢侈品的關係上呈現負相關，如magazine 的負相關值得係數最大。
- **Stores**
  - onpromotion 的商品越多，會帶給sales 更好影響
- **Holiday**
  - 整理完 holiday_events 後和 train data merge 後，
  - 經過 AB test 可留下需要考慮作為feature 的 holidays 和 events。

未來工作

選擇要輸入模型的features

資料合併及填補缺失值

建立模型並進行預測與評估

參考資料

- Store Sales TS Forecasting - A Comprehensive Guide | Kaggle
- Store Sales - Time Series Forecasting - Hyperparameters | Kaggle

Thank you for listening

Q&A