# Predict Future Sales

wei_ting-lin 2021/1/5

## (KDD0)數據來源

### 簡介

資料來源: Kaggle-Predict Future Sales

以2013.01~2015.10銷售數據預測2015.11每家商店賣出的產品數量

檔案說明:

- sales_train.csv-2013年1月至2015年10月的每日曆史數據
- test.csv-需要預測這些商店和產品在2015年11月的銷售額
- items.csv-項目/產品補充訊息
- item_categories.csv- 項目/類別補充訊息
- shop.csv-店家補充訊息

## (KDD1)數據擷取

### 讀入資料

訓練數據:sales_train.csv(sales):

sales.shape= (2935849, 6)

約290萬筆交易數據

sales.head(5)=

| | date | date_block_num | shop_id | item_id | item_price | item_cnt_day |
|---|---|---|---|---|---|---|
| 0 | 02.01.2013 | 0 | 59 | 22154 | 999 | 1 |
| 1 | 03.01.2013 | 0 | 25 | 2552 | 899 | 1 |
| 2 | 05.01.2013 | 0 | 25 | 2552 | 899 | -1 |
| 3 | 06.01.2013 | 0 | 25 | 2554 | 1,709.0500 | 1 |
| 4 | 15.01.2013 | 0 | 25 | 2555 | 1099 | 1 |

- date:商品售出日期
- date_block_num:連續月份(2013年1月為0)
- shop_id:此商品售出的店家
- item_id:此商品ID
- item_price:此商品標價
- item_cnt_day-當日銷售的產品數量

品項數據:items.csv(items):

items.shape= (22170, 3)

共有22170種品項

items.head(5)=

| | item_name | item_id | item_category_id |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 0 | ! ВО ВЛАСТИ НАВАЖДЕНИЯ… | 0 | 40 |
| 1 | !ABBYY FineReader 12 P… | 1 | 76 |
| 2 | ***В ЛУЧАХ СЛАВЫ (UNV)… | 2 | 40 |
| 3 | ***ГОЛУБАЯ ВОЛНА (Univ… | 3 | 40 |
| 4 | ***КОРОБКА (СТЕКЛО) D | 4 | 40 |

- item_name:品項名稱
- item_id:品項編碼(流水碼)
- item_category_id:品項對應類別

類別數據:item_categories.csv(item_cats):

item_cats.shape= (84, 2)

共有84種類別

item_cats.head(5)=

| | item_category_name | item_category_id |
|---|---|---|
| 0 | PC - Гарнитуры/Наушники | 0 |
| 1 | Аксессуары - PS2 | 1 |
| 2 | Аксессуары - PS3 | 2 |
| 3 | Аксессуары - PS4 | 3 |
| 4 | Аксессуары - PSP | 4 |

- item_category_name:類別名稱
- item_category_id:類別編碼(流水碼)

通路數據:shops.csv(shops):

shops.shape= (60, 2)

共有60家通路

shops.head(5)=

| | shop_name | shop_id |
|---|---|---|
| 0 | !Якутск Орджоникидзе, 56 фран | 0 |
| 1 | !Якутск ТЦ "Центральный" фран | 1 |
| 2 | Адыгея ТЦ "Мега" | 2 |
| 3 | Балашиха ТРК "Октябрь-Киномир" | 3 |
| 4 | Волжский ТЦ "Волга Молл" | 4 |

- shops_name:通路名稱
- shops_id:通路編碼(流水碼)

預測數據:test.csv(test):

test.shape= (60, 2)

共有214200個預測目標

test.head(5)=

| | ID | shop_id | item_id |
|---|---|---|---|
| 0 | 0 | 5 | 5037 |
| 1 | 1 | 5 | 5320 |
| 2 | 2 | 5 | 5233 |
| 3 | 3 | 5 | 5232 |
| 4 | 4 | 5 | 5268 |

- ID:流水碼
- shops_id:通路編碼

- item_id:品項編碼

# (KDD2)數據探索

## (2-1)創建年/月欄位

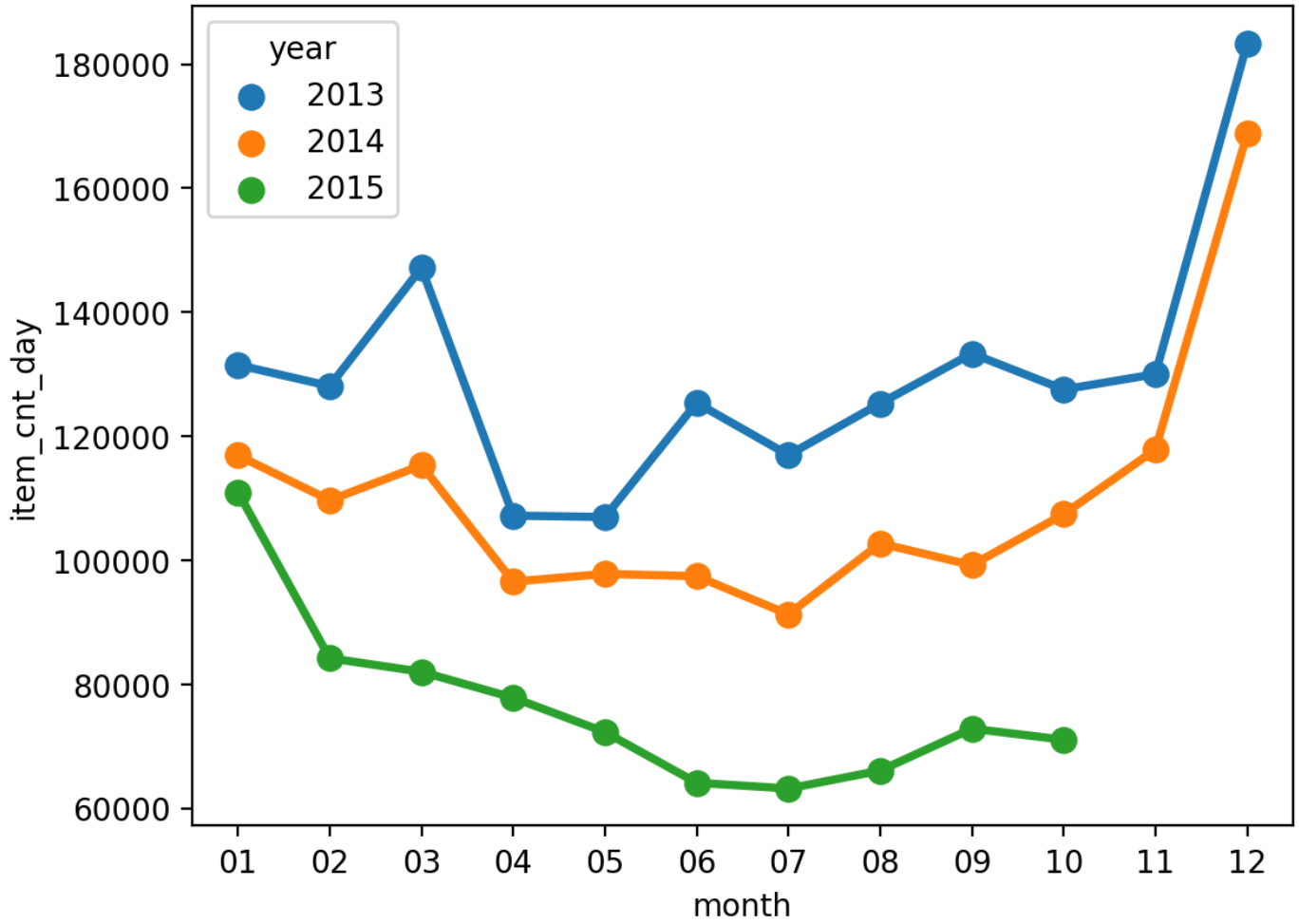
因為要預測的銷售數量是以月為單位，所以獨立出來以利分析

sales.head(5)=

| | date | date_block_num | shop_id | item_id | item_price | item_cnt_day | yea |
|---|---|---|---|---|---|---|---|
| 0 | 02.01.2013 | 0 | 59 | 22154 | 999 | 1 | 20 |
| 1 | 03.01.2013 | 0 | 25 | 2552 | 899 | 1 | 20 |
| 2 | 05.01.2013 | 0 | 25 | 2552 | 899 | -1 | 20 |
| 3 | 06.01.2013 | 0 | 25 | 2554 | 1,709.0500 | 1 | 20 |
| 4 | 15.01.2013 | 0 | 25 | 2555 | 1099 | 1 | 20 |

## (2-2)數據分布

歷年銷售數據(銷售量/月):



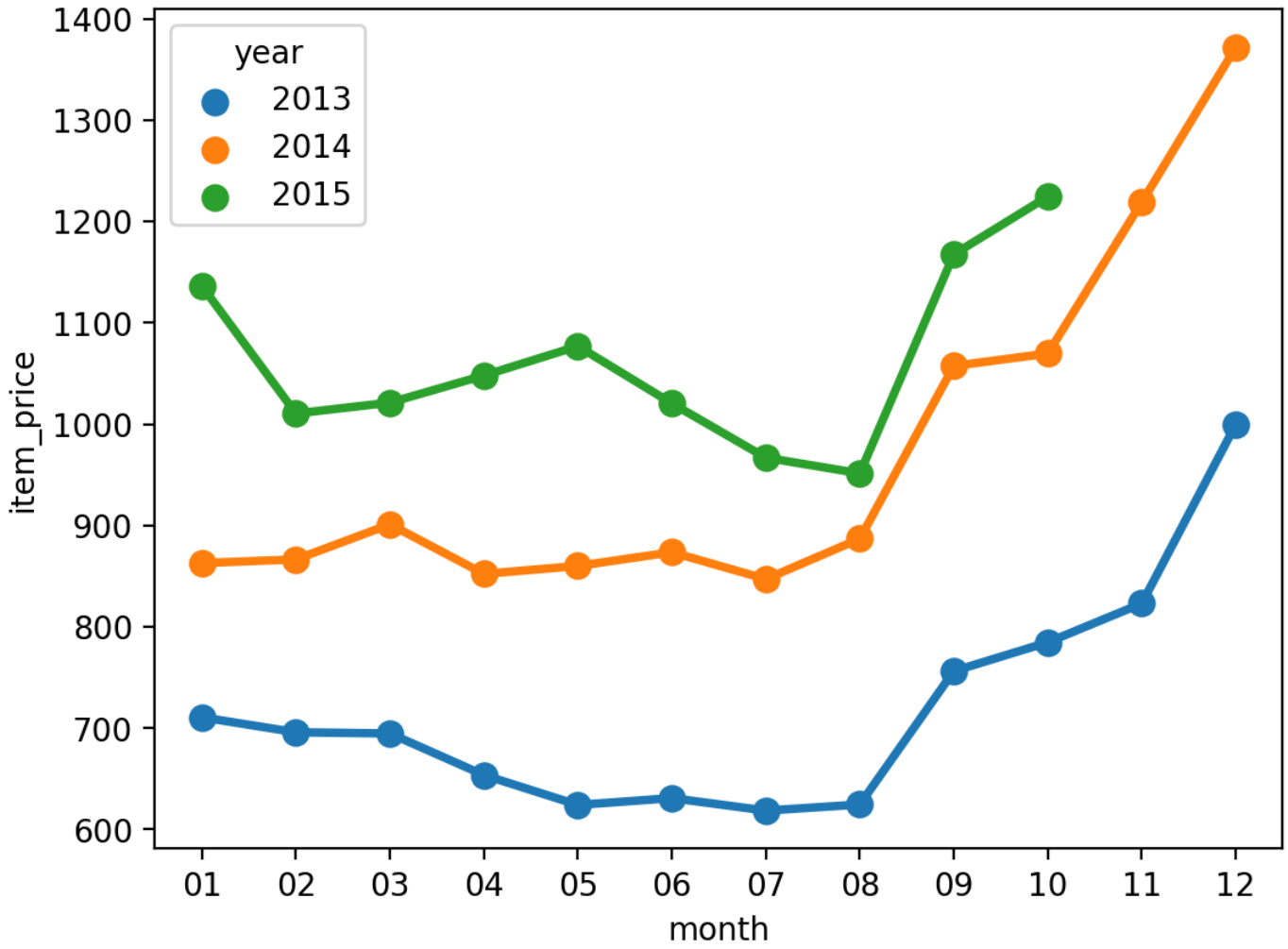

以整體銷售量數據來看

1.售出的商品數量有逐年減少的趨勢

2.在接近年尾時商品賣得較好

2.年尾是銷售量最高

歷年銷售數據(銷售金額/月):



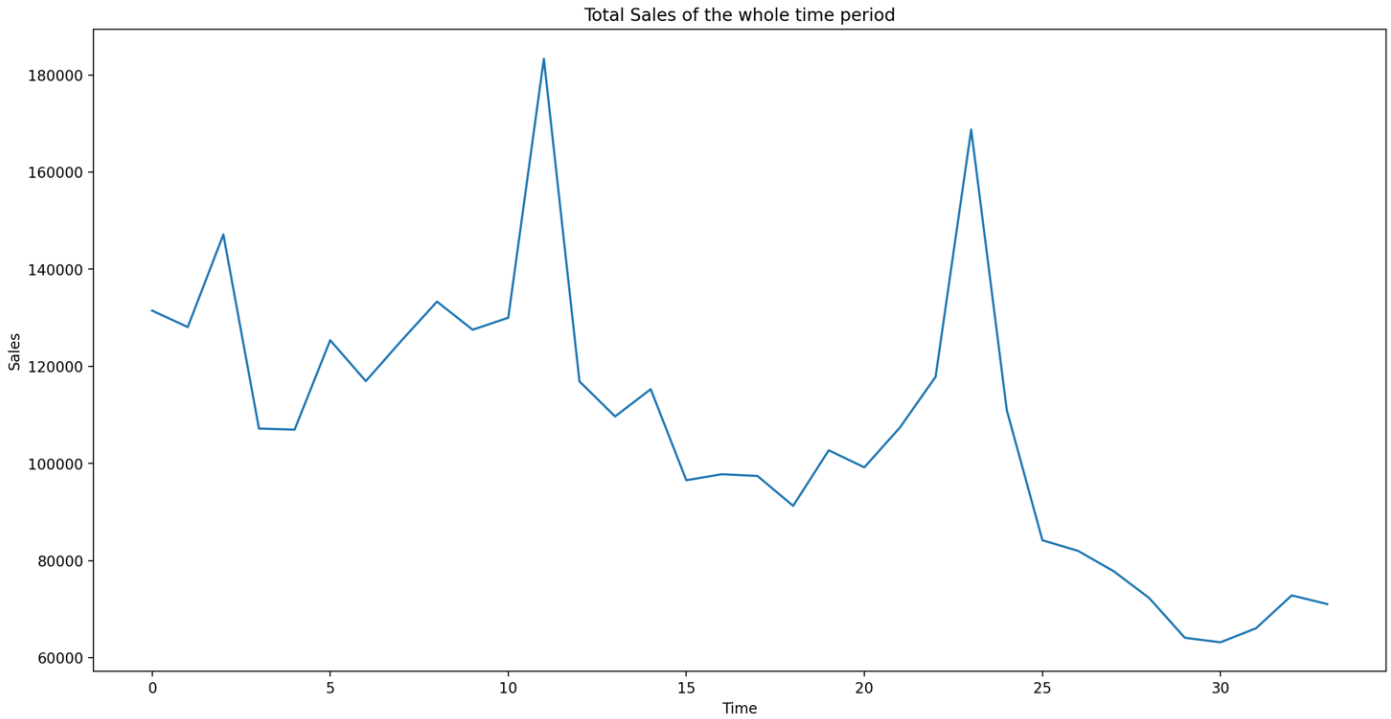

以整體銷售額數據來看

1.售出的商品價格逐年提高

2.每年8-9月金額增加幅度大

3.年尾是銷售額最高

連續月份銷量數據(銷售量/連續月份):



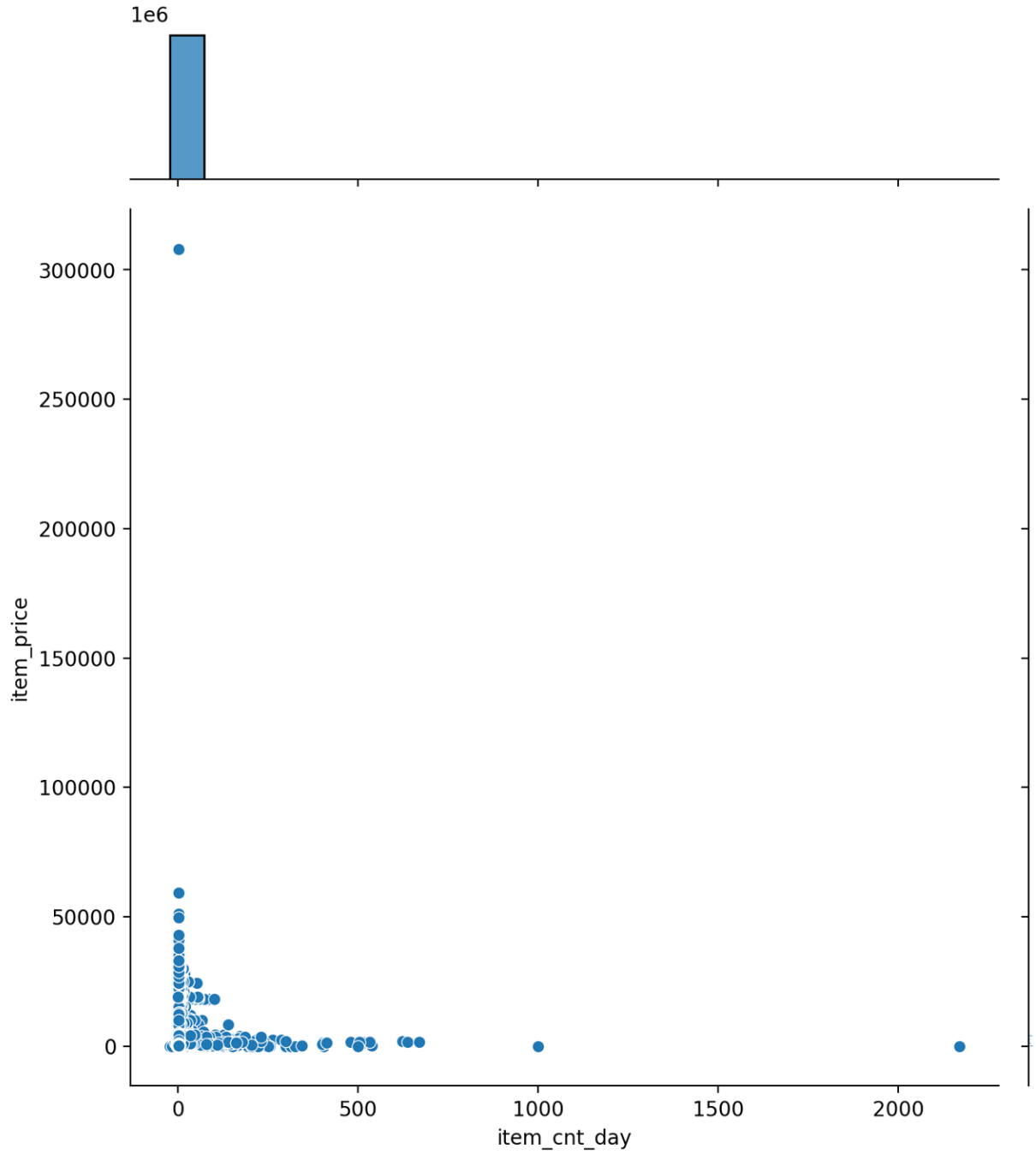

以連續月份銷售量來看

1.銷售量逐年下降

2.春夏之間銷量低迷

3.年底是銷量高峰

價格與銷量分布圖


1e6
300000
250000
200000
150000
100000
50000
0
item_price
0
500
1000
1500
2000
item_cnt_day

以價格與銷量分布圖來看

1.價格帶主要落在0~50000之間

2.單筆交易數量主要落在0~500之間

3.價格和單筆件數離異點

## (2-3)數據清理

數據清理前資料大小: `(2935849, 8)`

數據清理後資料大小: `(1221451, 8)`

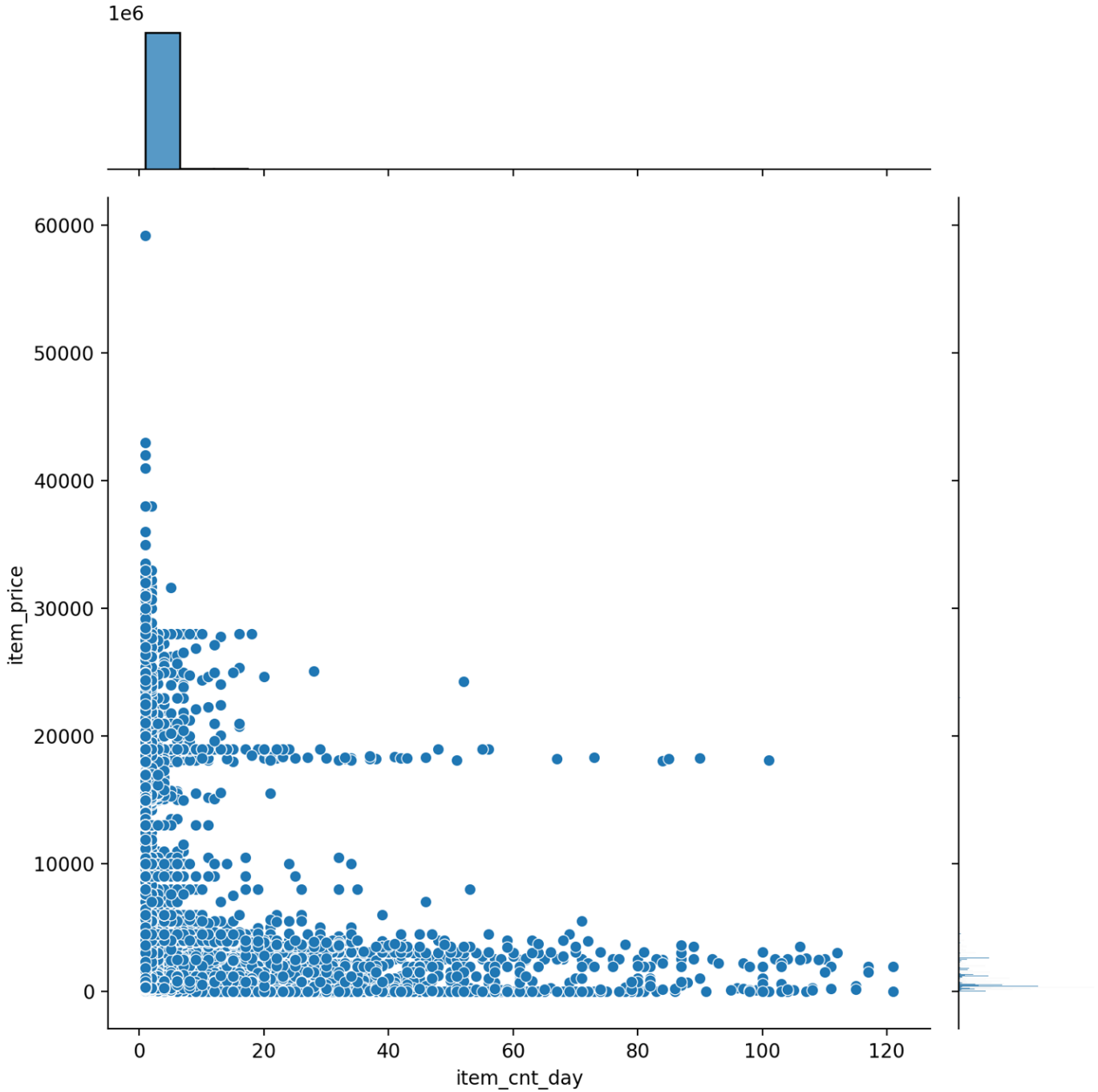
清除條件 :

1.單價小於0(退貨)

2.預測目標不包含的通路及品項

3.單筆件數大於125以及單筆金額大於75000(離異點)

清洗後的銷量價格分布:


1e6
60000
50000
40000
30000
20000
10000
0
item_price
0
20
40
60
80
100
120
item_cnt_day

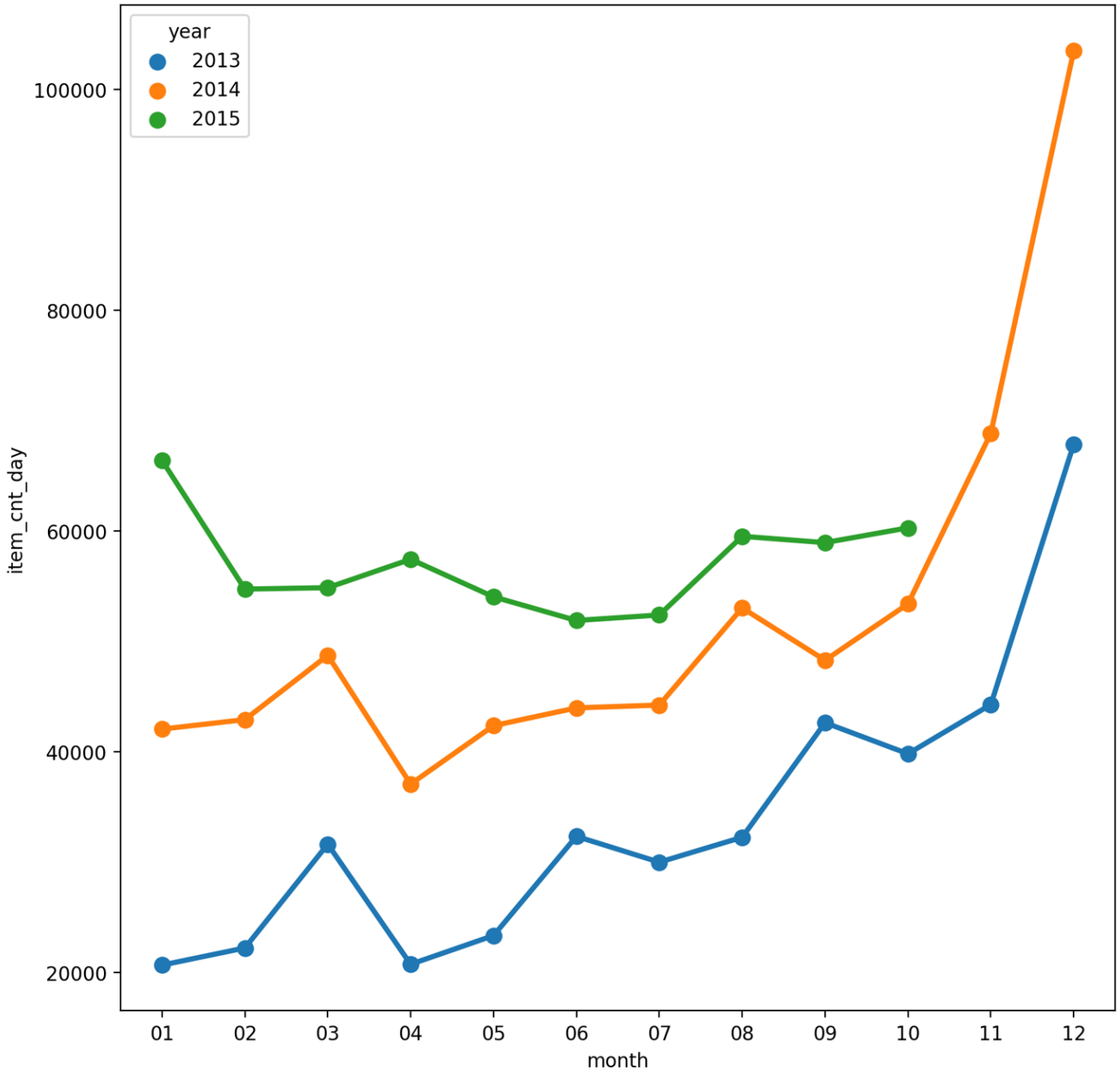
1.價格主要落在0~10000之間

2.單筆交易數量主要落在0~20之間

3.單價越低數量越高

4.單價20000附近的價格帶也有高銷量

資料清理後歷年銷售數據(數量/月):


year
2013
2014
2015
100000
80000
60000
40000
20000
item_cnt_day
01
02
03
04
05
06
07
08
09
10
11
12
month

以要預測的商品來看

1.是銷售數量逐年增加的產品

2.同樣的在接近年尾時商品賣得較好

小節

1.年尾銷售量銷售額都比較高

2.銷售量逐年下降

3.銷售額逐年上升

4.銷售量下降,銷售額卻上升 -> 價格提高 -> 俄羅斯金融風暴

# (KDD3)數據轉換

## 將**sales**數據依通路以及品項列出連續**33**個月的銷售量

pivoted_sales.head(5)=

| | sum | sum | sum | sum | sum | sum | sum | sum | sum | sum | sum | sum | sum | sum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ( |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ( |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ( |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ( |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ( |

將數據轉換成以通路品項為行,33個連續月份銷量為列的資料格式

## 通路聚類

高單價通路:

| | shop_id | item_cnt_day | item_price | group | one |
|---|---|---|---|---|---|
| 2 | 2 | 30620 | 35,104,442.4319 | 0 | 1,146.4547 |
| 4 | 4 | 43942 | 35,335,390.7161 | 0 | 804.1371 |
| 5 | 5 | 42762 | 33,260,876.5028 | 0 | 777.8139 |
| 6 | 6 | 100489 | 74,558,314.7803 | 0 | 741.9550 |
| 7 | 7 | 67058 | 56,077,104.6932 | 0 | 836.2478 |
| 12 | 12 | 73478 | 50,805,586.2852 | 0 | 691.4394 |
| 14 | 14 | 46375 | 36,006,448.2134 | 0 | 776.4194 |
| 15 | 15 | 71201 | 55,597,585.5484 | 0 | 780.8540 |
| 16 | 16 | 61633 | 48,712,980.1416 | 0 | 790.3717 |
| 18 | 18 | 65486 | 62,516,425.5943 | 0 | 954.6533 |
| 19 | 19 | 73455 | 54,302,000.4734 | 0 | 739.2553 |

通路數量: 31

單價區間: 638.8328221363008 - 1146.4546842567954

單價平均: 798.51

銷量區間: 30620.0 - 100489.0

銷量平均: 61634.29

總銷售額區間: 33260876.502807736 - 74558314.78033832

總銷售額平均: 48633587.09

高銷量通路:

| | shop_id | item_cnt_day | item_price | group | one |
|---|---|---|---|---|---|
| 25 | 25 | 241920 | 155,557,554.0090 | 1 | 643.0124 |
| 27 | 27 | 136657 | 105,648,893.1094 | 1 | 773.0954 |
| 28 | 28 | 184557 | 125,294,703.5110 | 1 | 678.8943 |
| 31 | 31 | 310777 | 170,763,434.3264 | 1 | 549.4726 |
| 42 | 42 | 144934 | 101,551,476.9729 | 1 | 700.6739 |
| 54 | 54 | 185790 | 109,669,430.1904 | 1 | 590.2870 |
| 57 | 57 | 141107 | 91,315,213.0305 | 1 | 647.1345 |

通路數量: 7

單價區間: 549.472561760871 - 773.0953636429902

單價平均: 654.65

銷量區間: 136657.0 - 310777.0

銷量平均: 192248.86

總銷售額區間: 91315213.03045203 - 170763434.32635823

總銷售額平均: 122828672.16

一般通路:

| | shop_id | item_cnt_day | item_price | group | one |
|---|---|---|---|---|---|
| 0 | 0 | 11705 | 5553869 | 2 | 474.4869 |
| 1 | 1 | 6311 | 2926161 | 2 | 463.6604 |
| 3 | 3 | 28355 | 26,472,615.0796 | 2 | 933.6137 |
| 8 | 8 | 3595 | 2,226,272.7050 | 2 | 619.2692 |
| 9 | 9 | 15866 | 4,714,302.7408 | 2 | 297.1324 |
| 10 | 10 | 24523 | 16,442,844.2300 | 2 | 670.5070 |

| | | | | | |
|---|---|---|---|---|---|
| 11 | 11 | 572 | 479,842.4600 | 2 | 838.8854 |
| 13 | 13 | 19763 | 5,333,601.4800 | 2 | 269.8781 |
| 17 | 17 | 25838 | 23,521,670.1235 | 2 | 910.3518 |
| 20 | 20 | 5872 | 2,389,265.9271 | 2 | 406.8913 |
| 23 | 23 | 7705 | 5,023,141.8461 | 2 | 651.9328 |

通路數量: 22

單價區間: 269.8781298385876 - 1192.9154270555048

單價平均: 710.25

銷量區間: 330.0 - 63388.0

銷量平均: 17809.14

總銷售額區間: 356819.0 - 29419582.061186437

總銷售額平均: 11271303.07

以通路聚類來看

1.高銷量通路平均總銷售額較高

2.高單價通路雖然單價比一般通路高，但平均銷售數也比較一般通路高

3.高銷量通路平均總銷售額明顯較高

# (KDD4)模型訓練

## (4-1)訓練集與測試集

X_train:

X_train.shape= (214200, 33, 1)

y_train:

y_train.shape= (214200, 1)

X_test:

X_test.shape= (214200, 33, 1)

## (4-2)模型建立

LSTM模型

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 64)                16896
_________________________________________________________________
dropout (Dropout)            (None, 64)                0
_________________________________________________________________
dense (Dense)                (None, 1)                 65
=================================================================
Total params: 16,961
Trainable params: 16,961
Non-trainable params: 0
_________________________________________________________________
```

```
Epoch 1/10
53/53 [==============================] - 22s 392ms/step - loss: 5.2446 - mean_squared_error: 5.2446
Epoch 2/10
53/53 [==============================] - 21s 392ms/step - loss: 4.0548 - mean_squared_error: 4.0548
Epoch 3/10
53/53 [==============================] - 21s 399ms/step - loss: 4.9020 - mean_squared_error: 4.9020
Epoch 4/10
53/53 [==============================] - 21s 399ms/step - loss: 4.6201 - mean_squared_error: 4.6201
Epoch 5/10
53/53 [==============================] - 21s 392ms/step - loss: 5.9941 - mean_squared_error: 5.9941
Epoch 6/10
53/53 [==============================] - 21s 390ms/step - loss: 6.2624 - mean_squared_error: 6.2624
Epoch 7/10
53/53 [==============================] - 21s 391ms/step - loss: 3.1035 - mean_squared_error: 3.1035
Epoch 8/10
53/53 [==============================] - 21s 392ms/step - loss: 8.5109 - mean_squared_error: 8.5109
Epoch 9/10
53/53 [==============================] - 21s 391ms/step - loss: 5.7937 - mean_squared_error: 5.7937
Epoch 10/10
53/53 [==============================] - 21s 392ms/step - loss: 4.2692 - mean_squared_error: 4.2692
```

回歸模型

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_4 (Dense)              (None, 64)                2176
_________________________________________________________________
dense_5 (Dense)              (None, 64)                4160
_________________________________________________________________
dense_6 (Dense)              (None, 1)                 65
=================================================================
Total params: 6,401
Trainable params: 6,401
Non-trainable params: 0
_________________________________________________________________
```

```
Epoch 1/10
53/53 [==============================] - 4s 72ms/step - loss: 8.1605 - mean_squared_error: 8.1605
Epoch 2/10
53/53 [==============================] - 4s 71ms/step - loss: 3.2143 - mean_squared_error: 3.2143
Epoch 3/10
53/53 [==============================] - 4s 71ms/step - loss: 3.8477 - mean_squared_error: 3.8477
Epoch 4/10
53/53 [==============================] - 4s 72ms/step - loss: 3.8677 - mean_squared_error: 3.8677
Epoch 5/10
53/53 [==============================] - 4s 71ms/step - loss: 3.8501 - mean_squared_error: 3.8501
Epoch 6/10
53/53 [==============================] - 4s 71ms/step - loss: 4.1684 - mean_squared_error: 4.1684
Epoch 7/10
53/53 [==============================] - 4s 71ms/step - loss: 5.5452 - mean_squared_error: 5.5452
Epoch 8/10
53/53 [==============================] - 4s 71ms/step - loss: 4.0857 - mean_squared_error: 4.0857
Epoch 9/10
53/53 [==============================] - 4s 71ms/step - loss: 3.5143 - mean_squared_error: 3.5143
Epoch 10/10
53/53 [==============================] - 4s 71ms/step - loss: 3.9686 - mean_squared_error: 3.9686
```

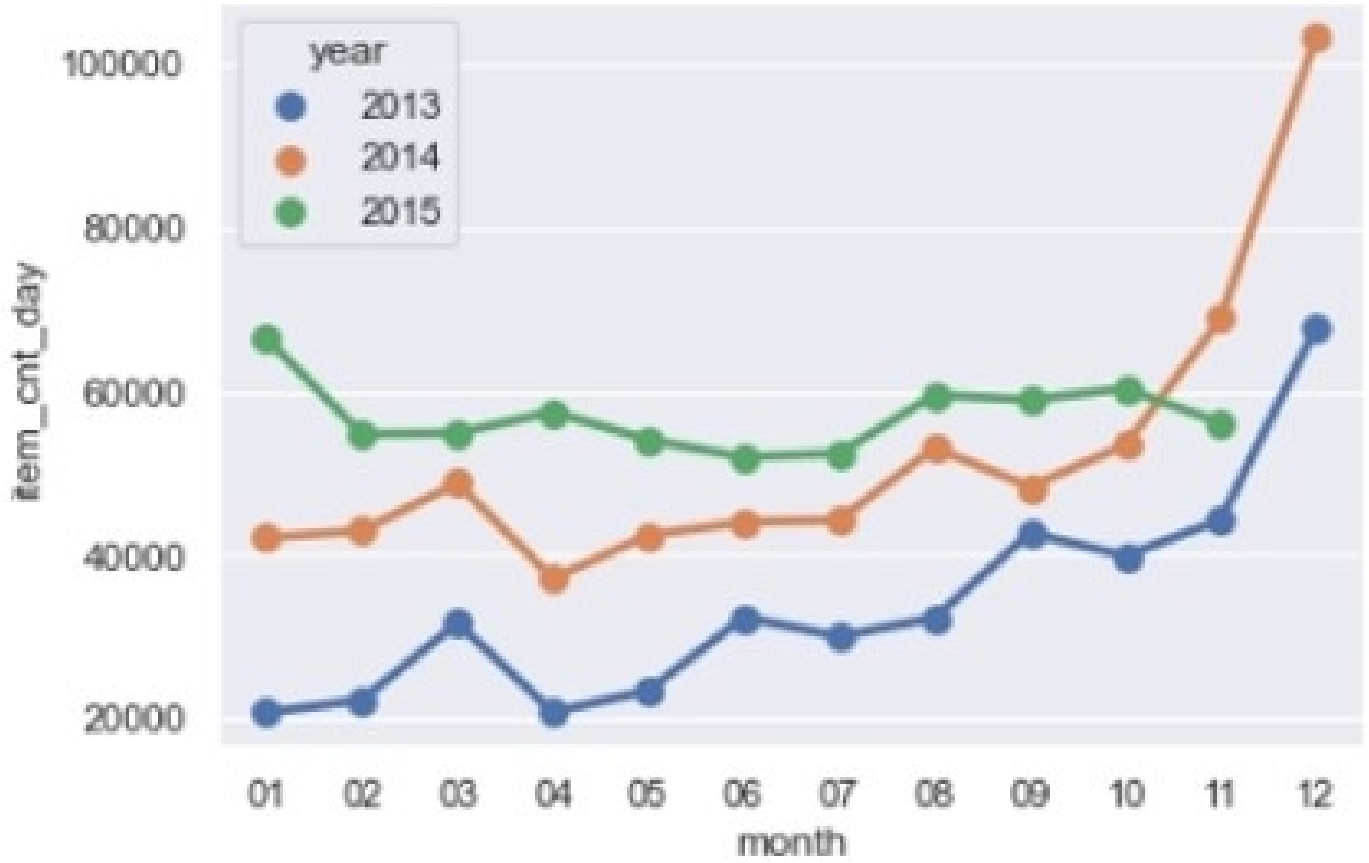
# (KDD5)預測

## 預測結果

submission1=

| | item_cnt_month |
|---|---|
| 0 | 0.5239 |
| 1 | 0.1131 |
| 2 | 0.8700 |
| 3 | 0.1757 |
| 4 | 0.1131 |

lstm預測後歷年銷售數據(數量/月):




Sunrise by the mountains

預測11月銷量些微下降

依歷年數據來看即便沒有大幅度提升也應該要有上升趨勢

# 生成word檔

** 已產生報告(sales.docx)

Made with Streamlit