Corporación Favorita Grocery Sales Forecasting

Business Content

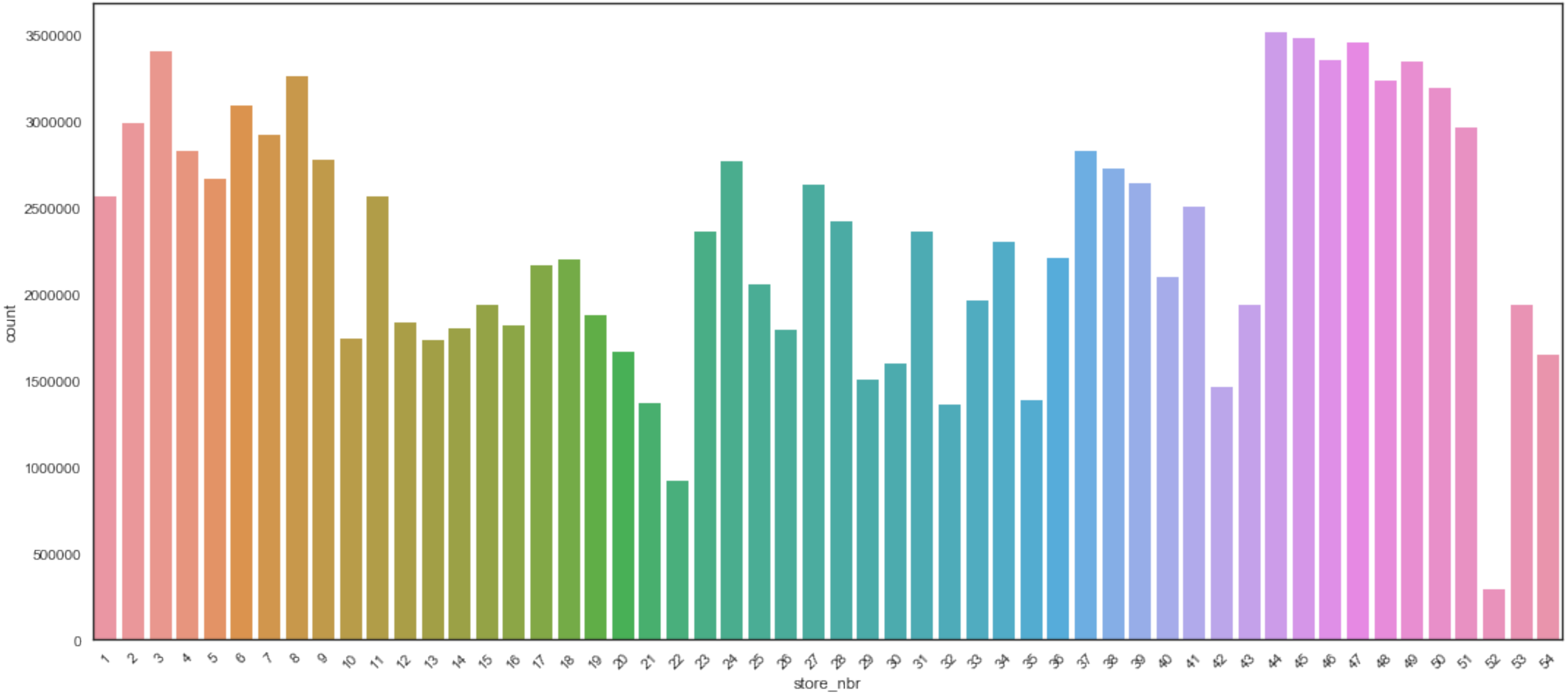
- The stock in the grocery is highly influenced by the predicted sales

- Corporación Favorita, a large Ecuadorian-based grocery retailer, tries to find a solution in machine learning models to accurately predict the unit sales for thousands of items sold at different Favorita stores located in Ecuador.

Data Description

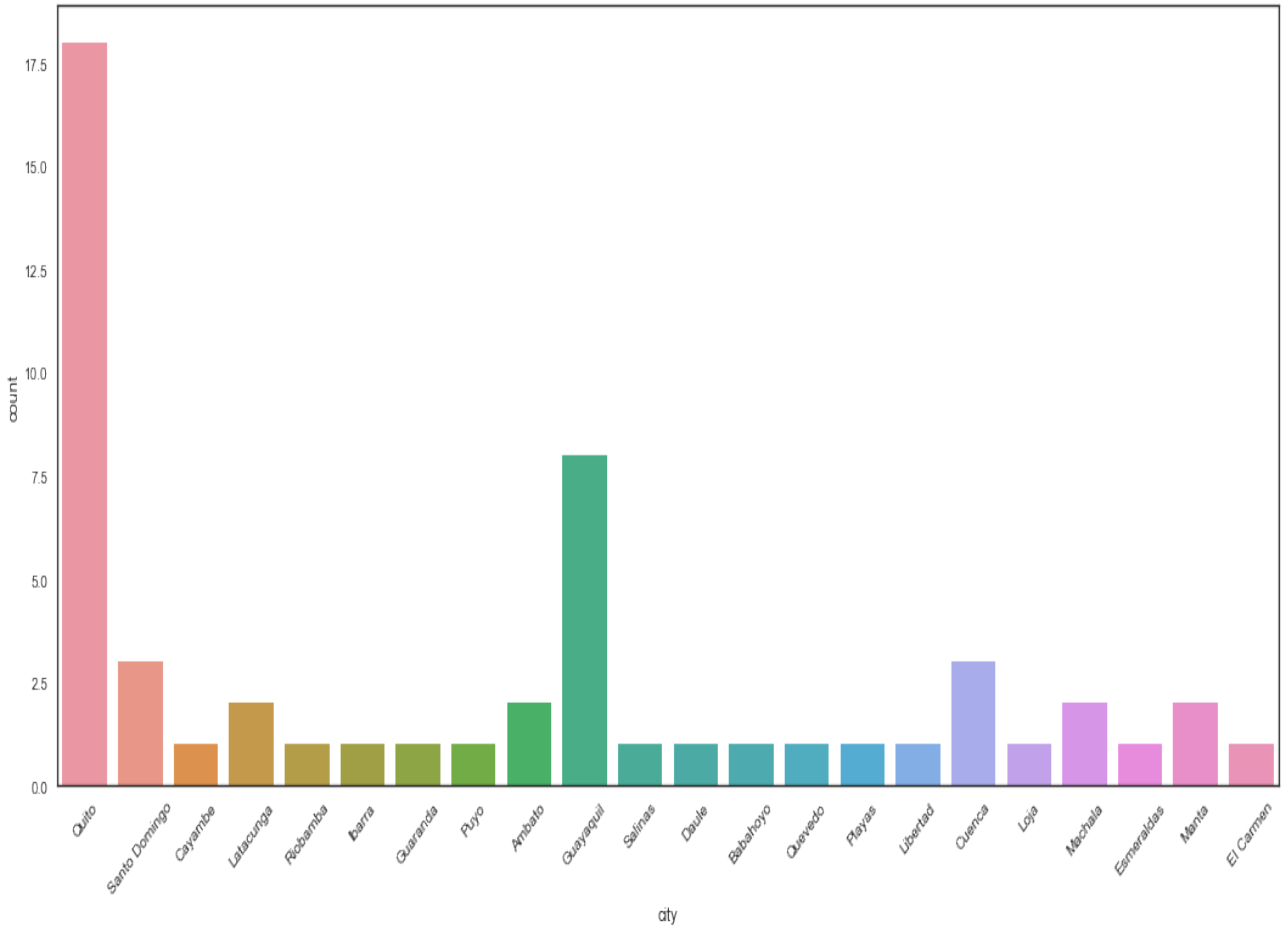
- Training data : the unit_sales and promotion state by date, store number and item number
- Store meta data : city, state, type and cluster (a grouping of similar stores) for particular store numbers
- Items meta data: family, class and perishable for particular item numbers
- Oil data: daily oil price
- Transaction data: daily transactions
- Holiday events data: holiday type and date

EDA_Store information

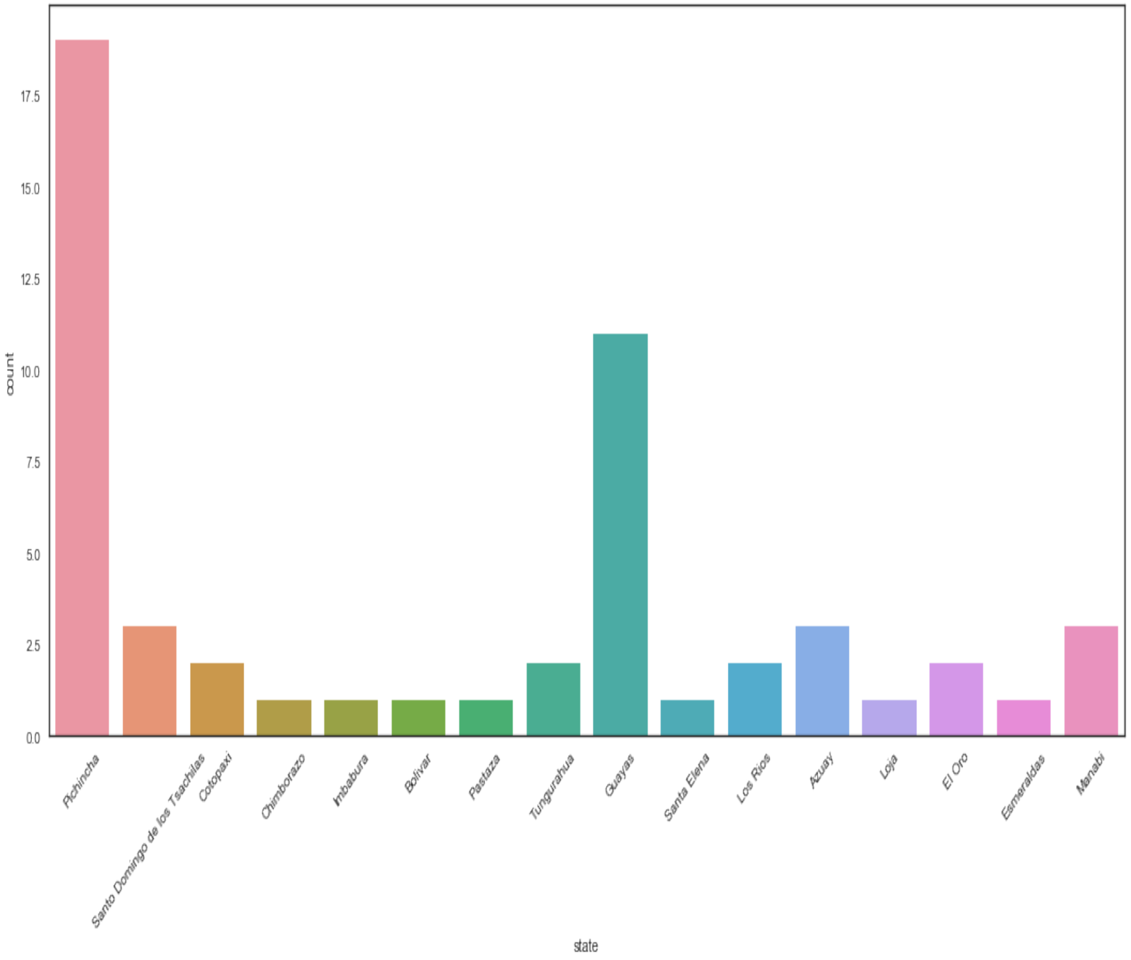


Majority of stores have items sale records more than 1,500,000 since January 2013.
An interesting store is number 52, only has less than 500,000 sale records.

EDA_Store information

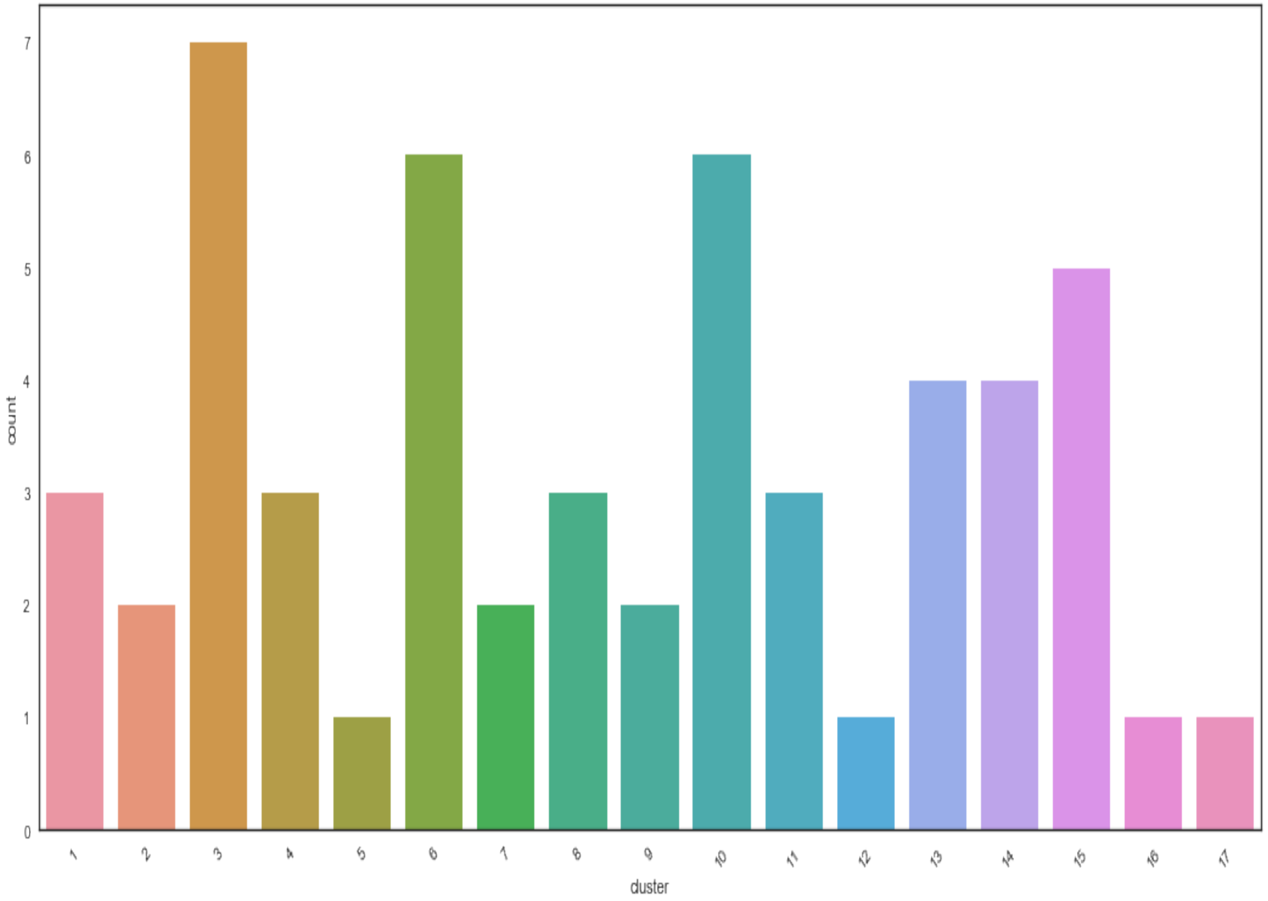


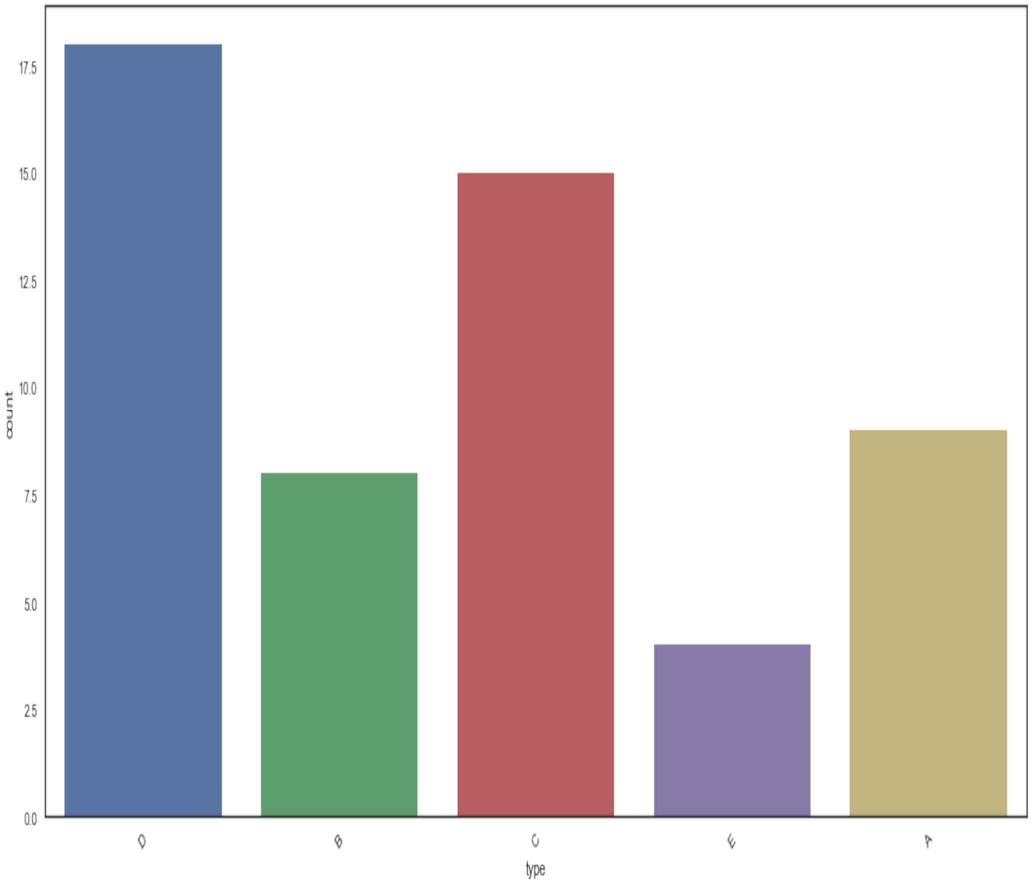
City of Quito and Guayaquil take the leading board



State Rchincha and Guayas outnumber other states

EDA_Store information

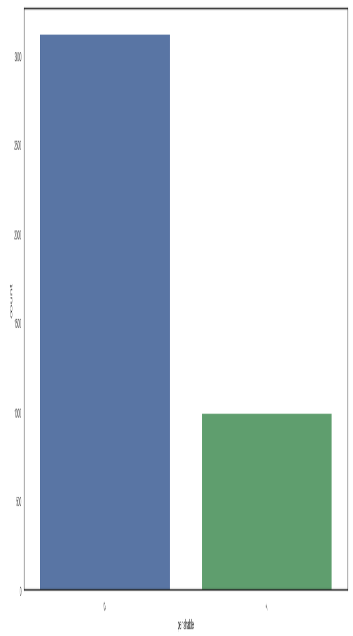


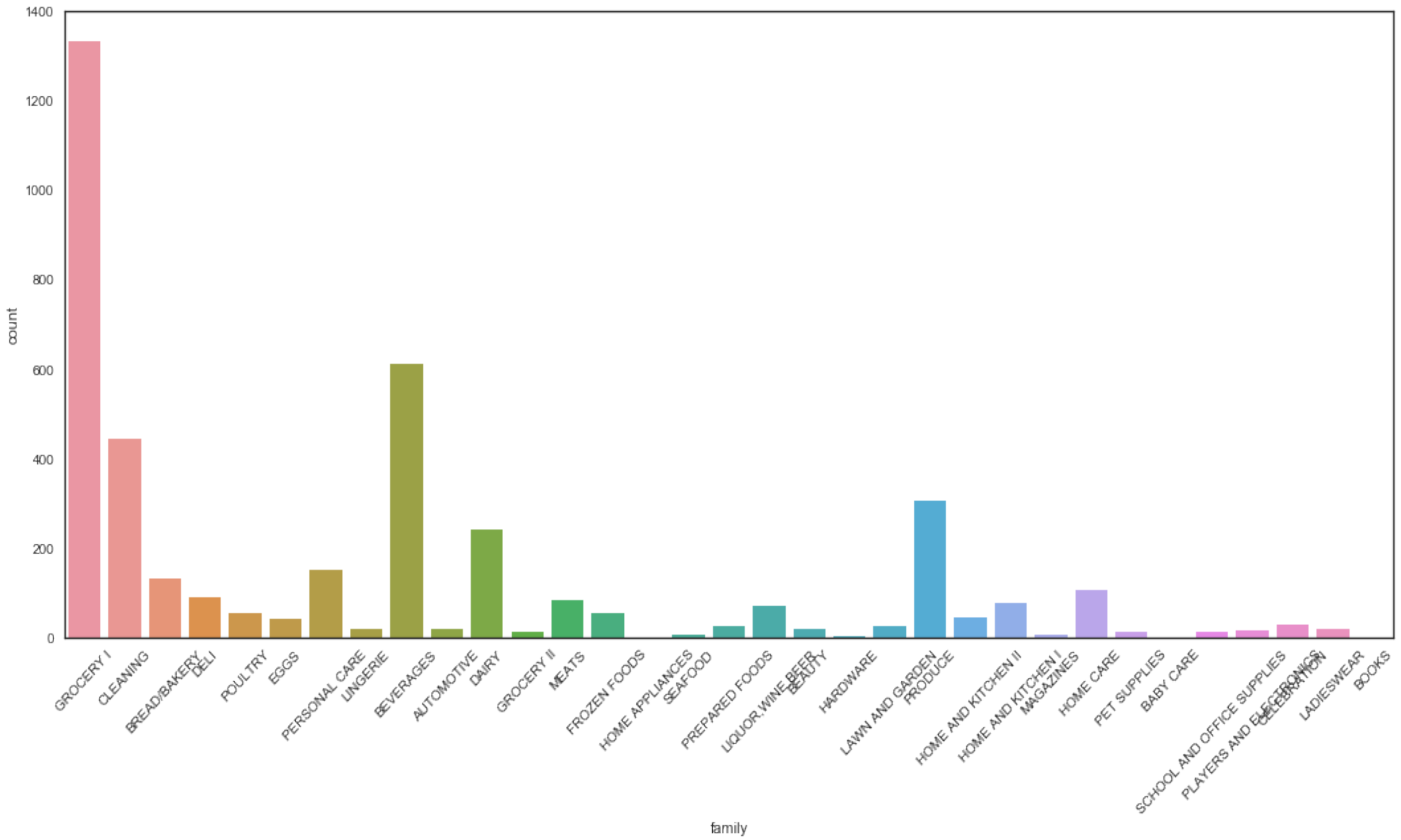


Difference of store numbers for different clusters is not big

Difference of store numbers for different types is not big

EDA_Item information

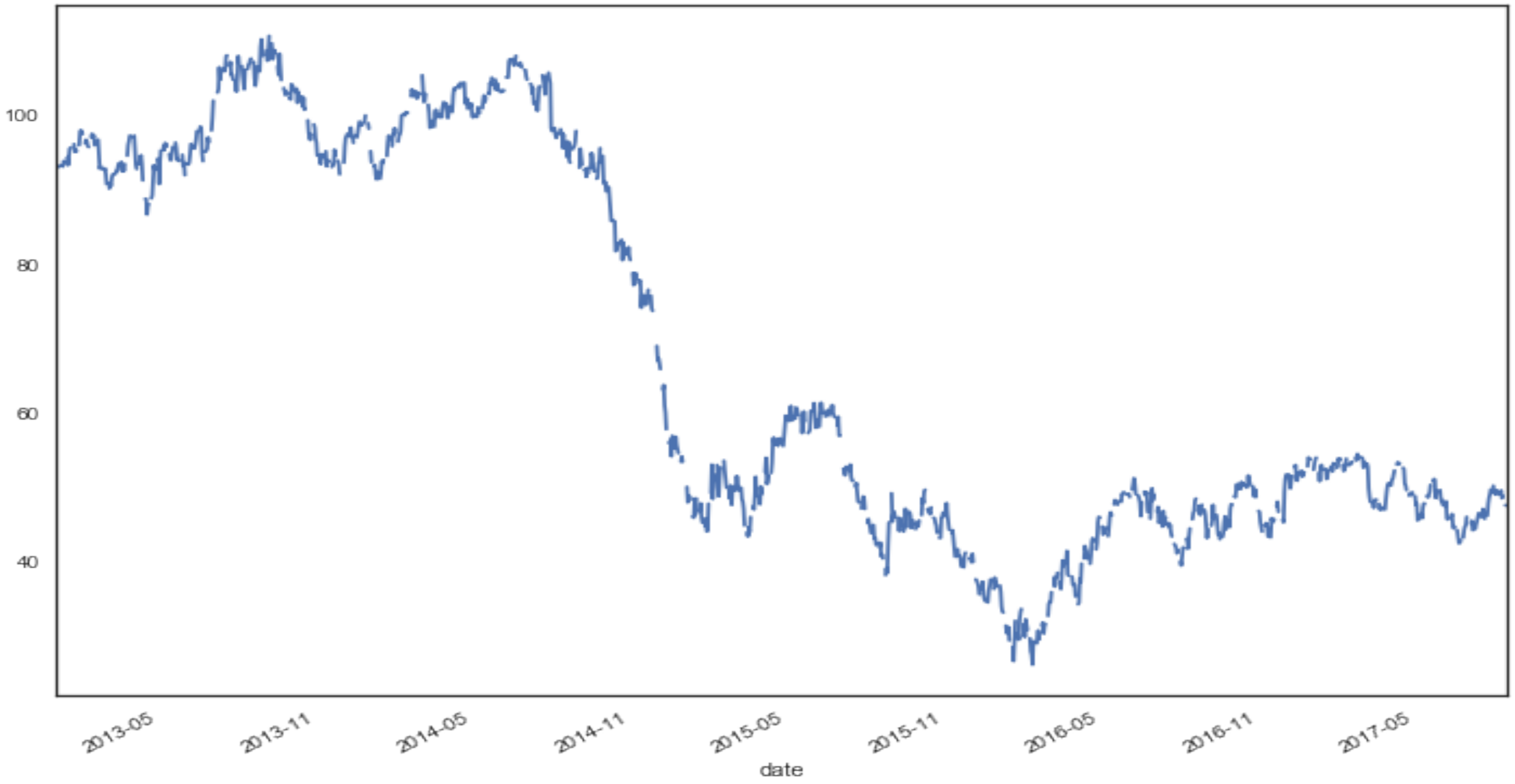




Items of non-perishable categories roughly 3 times compared with items in perishable ones

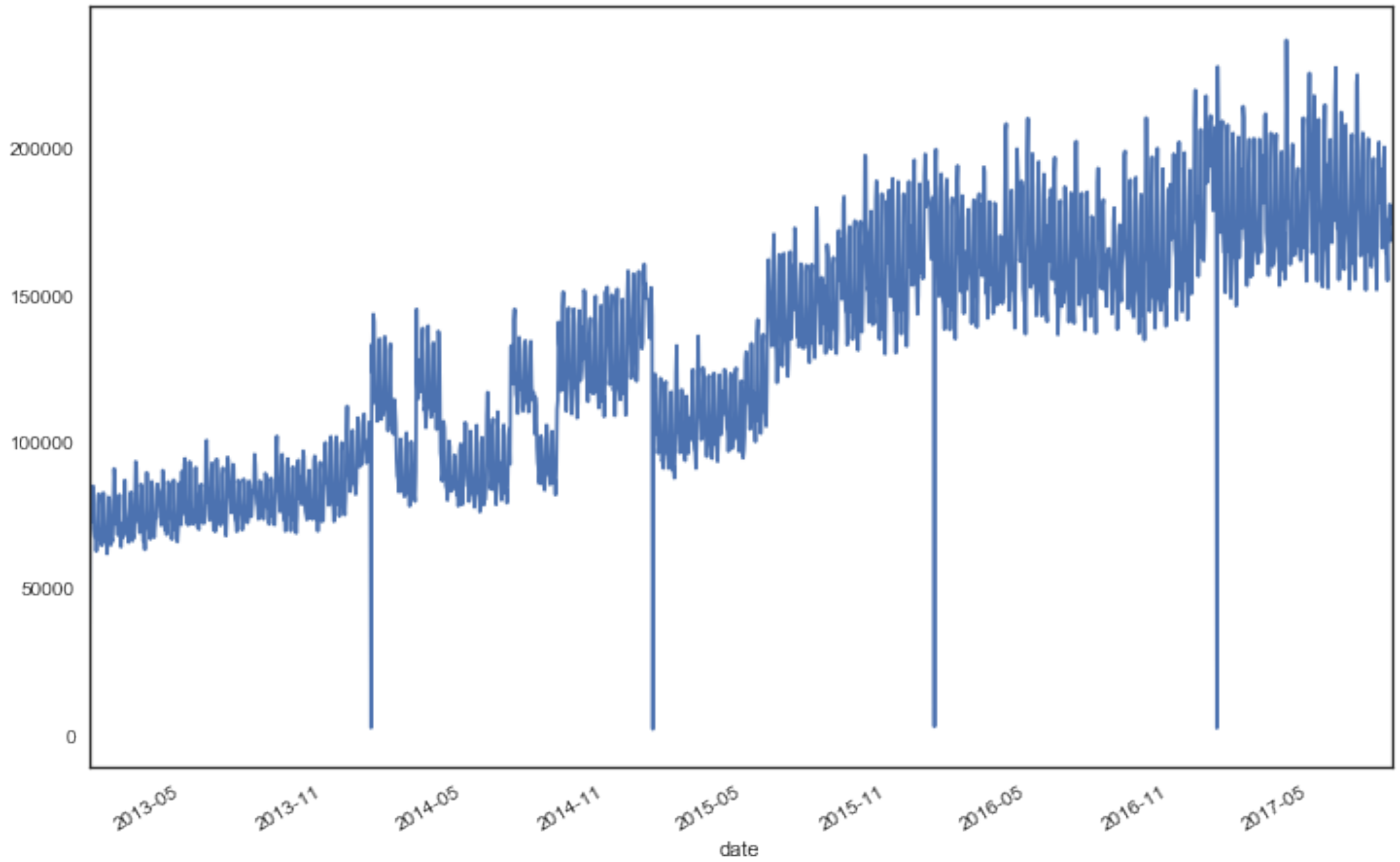
Majority of items fall into the grocery and beverage family

EDA_Oil price



A decline trend in the long terms
A huge drop is seen at the beginning of 2015
A growing oil price are both observed in the recent data since 2016

EDA_Total unit sales



An increasing trend in the long term
A fluctuation in the short term
Several significant drops happened around Christmas, which was closed of stores on those days.

Inferential statistics

- Is there any statistically relationship between oil price and total unit sales?

  Yes (liner regression p-value =0)

- Is promotion and unit sales independent each other?

  Yes (student t test p-value =0)

- Is store type and cluster independent with each other and is store city and cluster in dependent with each other

  Yes on store type and cluster and No on store city and cluster (chi-square test p-value =0 and 0.73)

Data Wrangling and Feature engineering

- Stage1. Feature transformation and selection

Loading data with unit_sale negitve values transformed by log(1+x)

Filling the missing values with forward filling in oil price and 0 for promotion state

National holiday date selection(no match in testing data)

Category label encoding

Mapping perishable and promotion features

Data Wrangling and Feature engineering

- Stage1. Feature transformation and selection

We prefer to learn from the recent past information

Choose the date from 2017 as a start point.

Training data set will be the major data source for building up model

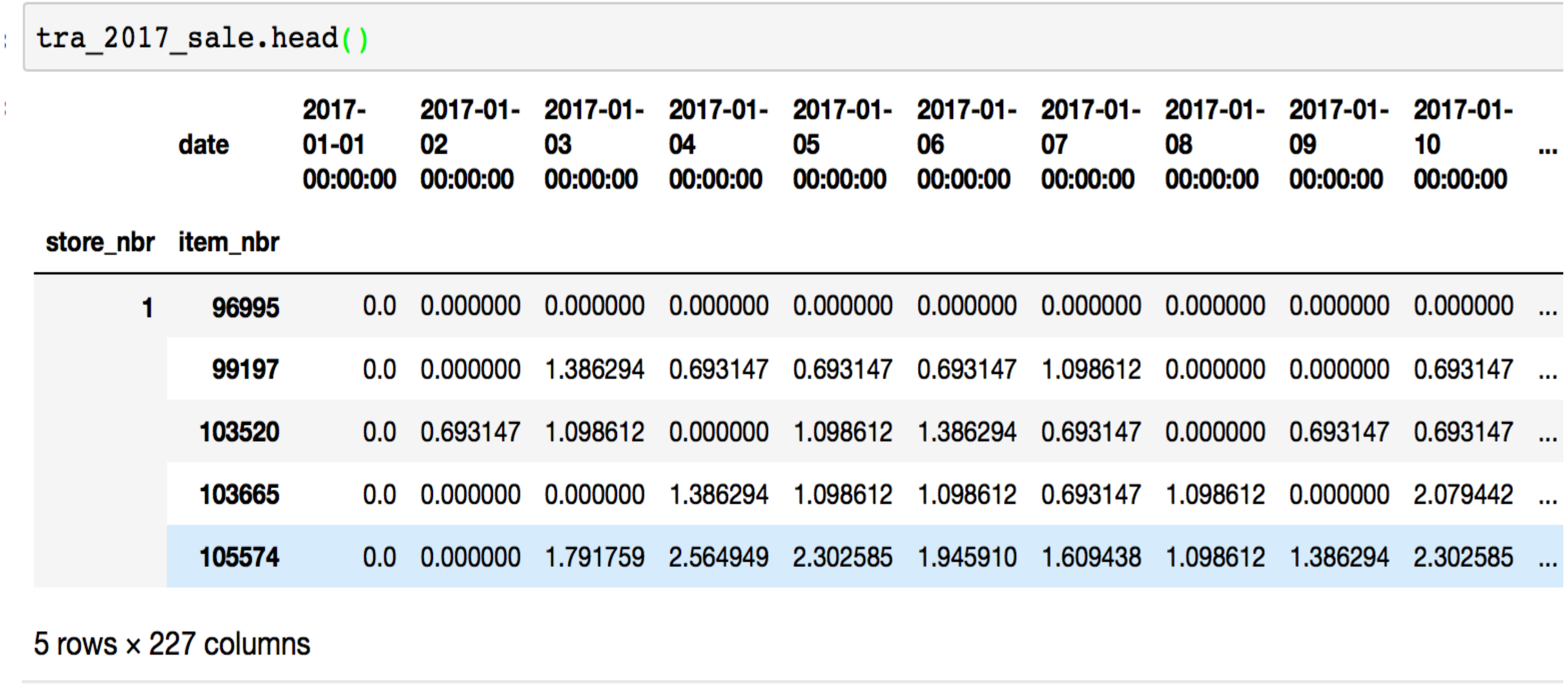
Oil data doesn’t provide strong evidence on influencing short term forecast

Holiday data indicates no big events happened in the prediction range

Store and item meta data give some information and we can use these data to create cross features to improve the predict ability of model

Stage2. Create Time series features

- Data Frame reshape



```
tra_2017_sale.head()
```

	date	2017-01-01 00:00:00	2017-01-02 00:00:00	2017-01-03 00:00:00	2017-01-04 00:00:00	2017-01-05 00:00:00	2017-01-06 00:00:00	2017-01-07 00:00:00	2017-01-08 00:00:00	2017-01-09 00:00:00	2017-01-10 00:00:00	...
store_nbr	**item_nbr**											
1	**96995**	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
	99197	0.0	0.000000	1.386294	0.693147	0.693147	0.693147	1.098612	0.000000	0.000000	0.693147	...
	103520	0.0	0.693147	1.098612	0.000000	1.098612	1.386294	0.693147	0.000000	0.693147	0.693147	...
	103665	0.0	0.000000	0.000000	1.386294	1.098612	1.098612	0.693147	1.098612	0.000000	2.079442	...
	105574	0.0	0.000000	1.791759	2.564949	2.302585	1.945910	1.609438	1.098612	1.386294	2.302585	...

5 rows × 227 columns

Stage2. Create Time series features

- Define time window functions

Moving statistics for days of interval: mean, mean of difference for adjacent, decay mean, median, min, max, standard deviation

Moving aggregation for days of interval: sum

Select values of consecutive days

Stage2. Create Time series features

- Create time series features based on the defined time window functions

Input features (X)

Unti_sale time series data

Moving statistics of unit_sale in 3, 7, 14, 30, 60, 140 past days of interval

Moving aggregation of unit_sale in 3, 7, 14, 30, 60, 140 past days of interval

Unit_sale in 16 past consecutive days (where 16 is the prediction date range)

Promotion state time series data

Moving aggregation of promotion state in 14, 60, 140 past days of interval

Moving aggregation of promotion state in 3, 7, 14 future days of interval

Promotion state in 16 past consecutive

Promotion state in 16 future consecutive days

Labels (y)

Select unit_sale in 16 future consecutive days

Machine Learning models

- Prepare training, validation and test dataset

Training date set : Four folders on 2017-06-14, 2017-06-21, 2017-06-28, 2017-07-05; then concatenate them together

Validate date set: 2017-07-26

Test data set : 2017-08-16.

Machine Learning models

Each of model select one day’s unit sales as label for training

Then predict correspondent unit sales in the 16 future consecutives days

Model	Performance/Error
Liner regression models_Lasso	0.3370012403947538
Liner regression models_Ridge	0.32294528746799755
Support Vector Machine	0.32922112263898434
Extra Trees	0.33125629069949236
Random Forest	0.33167082556360583

Machine Learning models

- 16 Light GBM models for training and prediction

Train 16 light GBM models : fast and more accurate

Performance error on validation data set is 0.314

Error in the test dataset is 0.515;

Ranked within 10% in the Kaggle leader board.