FLIP(00) Mid-term Presentation

Rongxin Xu
Hunan University

26 October 2019

Outline

Problem Statement

Exploratory Data Analysis

Exploratory Data Analysis

Stationarity

Conclusion

Problem Statement

Problem Description

This is a problem with time-series prediction. After a month of making scientific observations and taking careful measurements, can predict total sales for every product and store in the next month. The raw dataset contains train set with 2935849 samples and 214200 unlabeled samples as test set. Through the train data, predict total sales for every product and store in the next month.

Data Set

Defn There are 6 data sets with a total of 11 attributes, the fllowings are the name and meaning of attributes.

Data List

id an Id that represents a (Shop, Item) tuple within the test set.
shop_id unique identifier of a shop.
item_id unique identifier of a product.
item_category_id unique identifier of item category.
item_cnt_day percentage of soul in the creature.
item_price current price of an item.
date date in format dd/mm/yyyy.
date_block_num unique identifier of item category.
item_name name of item.
shop_name name of shop.
item_category_name name of item category.

Exploratory Data Analysis

Data Information

The following is the statistical result of each attribute in sales_train.csv. There are 6 numerical variables, and no missing values. The data is very clean and complete, So let's start visual analysis.

	date_block_num	shop_id	item_id	item_price	item_cnt_day	item_category_id
count	2935849	2935849	2935849	2935849	2935849	2935849
mean	14.57	33	10197.23	890.62	1.24	40
std	9.42	16.23	6324.3	1726.44	2.62	17.1
min	0	0	0	-1	-22	0
25%	7	22	4476	249	1	28
50%	14	31	9343	399	1	40
75%	23	47	15684	999	1	55
max	33	59	22169	307980	2169	83

Detailed description

	date	date_block_num	shop_id	item_id	item_price	item_cnt_day
0	02.01.2013	0	59	22154	999.00	1.0
1	03.01.2013	0	25	2552	899.00	1.0
2	05.01.2013	0	25	2552	899.00	-1.0
3	06.01.2013	0	25	2554	1709.05	1.0
4	15.01.2013	0	25	2555	1099.00	1.0
5	10.01.2013	0	25	2564	349.00	1.0

(a) sales_train.csv

	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
5	Аксессуары - PSVita	5

(b) item_categories.csv

	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
5	5	5	5039

(c) test.csv

	item_name	item_id	item_category_id
0	! ВО ВЛАСТИ НАВАЖДЕНИЯ (ПЛАСТ.) D	0	40
1	!ABBYY FineReader 12 Professional Edition Full...	1	76
2	***В ЛУЧАХ СЛАВЫ (UNV) D	2	40
3	***ГОЛУБАЯ ВОЛНА (Univ) D	3	40
4	***КОРОБКА (СТЕКЛО) D	4	40
5	***НОВЫЕ АМЕРИКАНСКИЕ ГРАФФИТИ (UNI) ...	5	40

(d) items.csv

	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
5	Вологда ТРЦ "Мармелад"	5

(e) shops.csv

Figure 1: Data Description

Summary

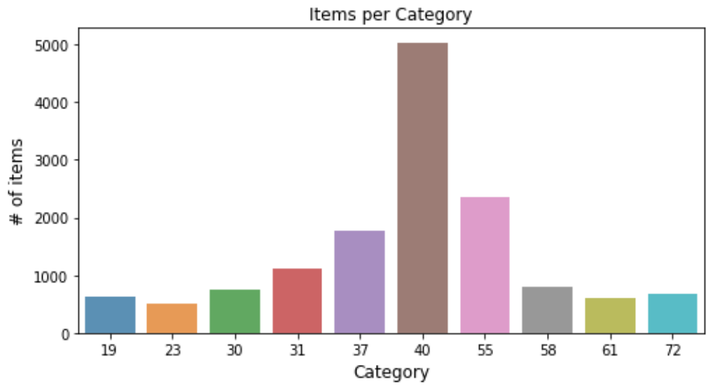
- The data is very clean and complete, so we only need to change the data type after importing.
- We also need to reorganize the table structure to make it more readable. An sample is given below.

			date		item_price	item_cnt_day
			min	max	mean	sum
date_block_num	shop_id	item_id				
0	0	32	2013-01-03	2013-01-31	221.0	6.0
		33	2013-01-03	2013-01-28	347.0	3.0
		35	2013-01-31	2013-01-31	247.0	1.0
		43	2013-01-31	2013-01-31	221.0	1.0
		51	2013-01-13	2013-01-31	128.5	2.0
		61	2013-01-10	2013-01-10	195.0	1.0
		75	2013-01-17	2013-01-17	76.0	1.0
		88	2013-01-16	2013-01-16	76.0	1.0
		95	2013-01-06	2013-01-06	193.0	1.0

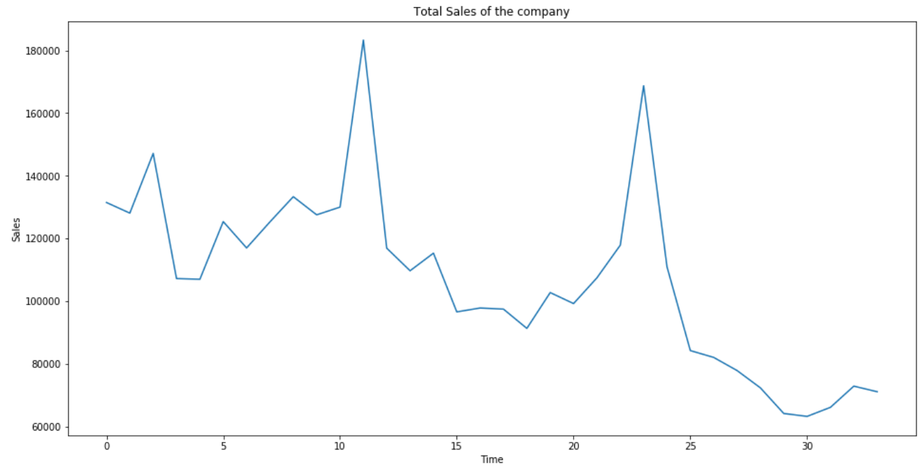
Figure 2: sample

Exploratory Data Analysis

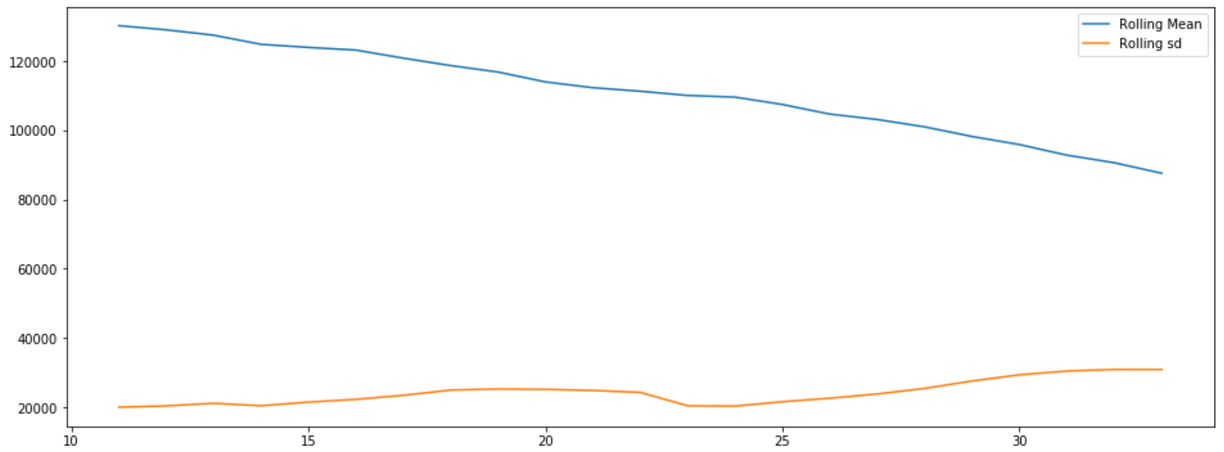
Exploratory Data Analysis



(a) Items per Category



(b) Total Sales of the company



(c) Rolling Mean and std

Figure 3: EDA

Summary

- There is an obvious "seasonality" (Eg: peak sales around a time of year) and a decreasing "Trend".

Stationarity

Seasonality and Trend

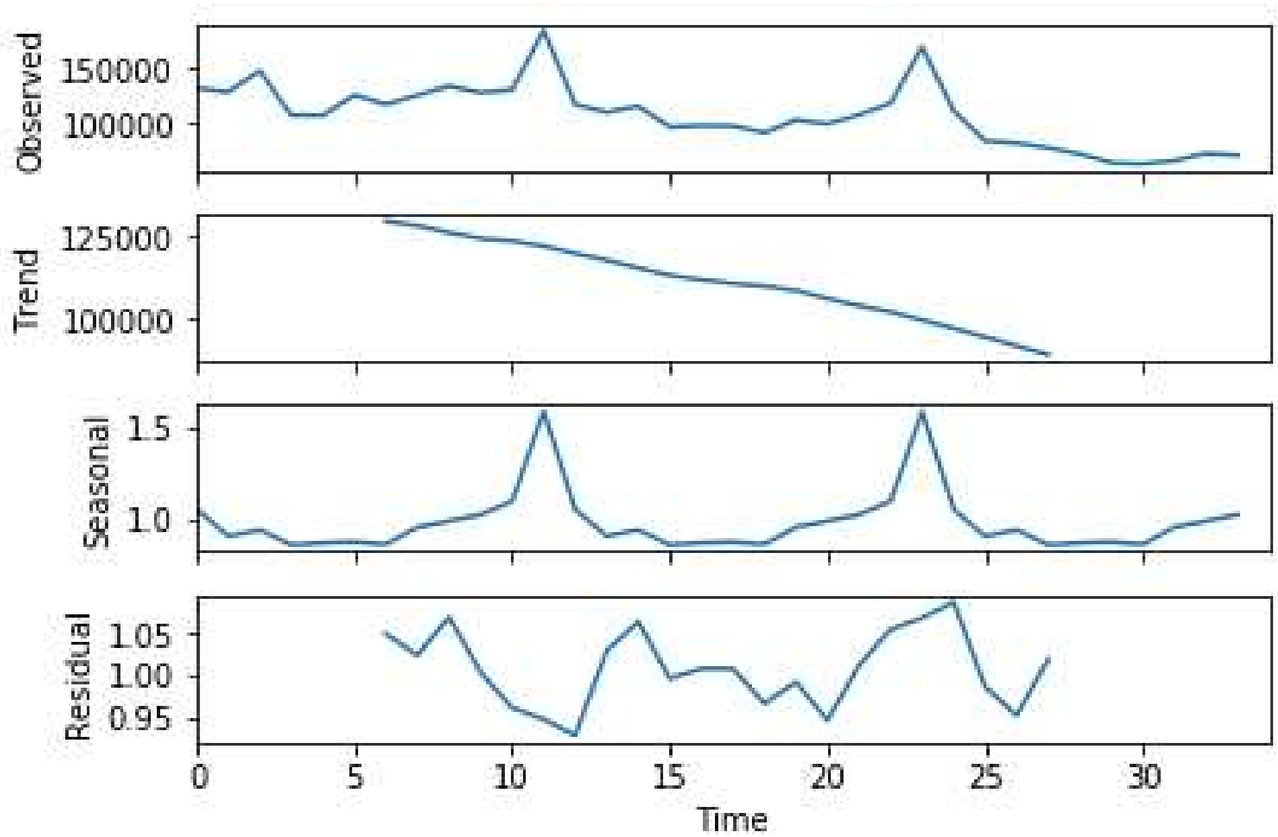


Figure 4: Seasonality and Trend

Stationarity Test


```
Results of Dickey-Fuller Test:
Test Statistic                  -2.395704
p-value                          0.142953
#Lags Used                       0.000000
Number of Observations Used     33.000000
Critical Value (1%)             -3.646135
Critical Value (5%)             -2.954127
Critical Value (10%)            -2.615968
dtype: float64
```

Figure 5: Stationarity Test

Remove seasonality and trends

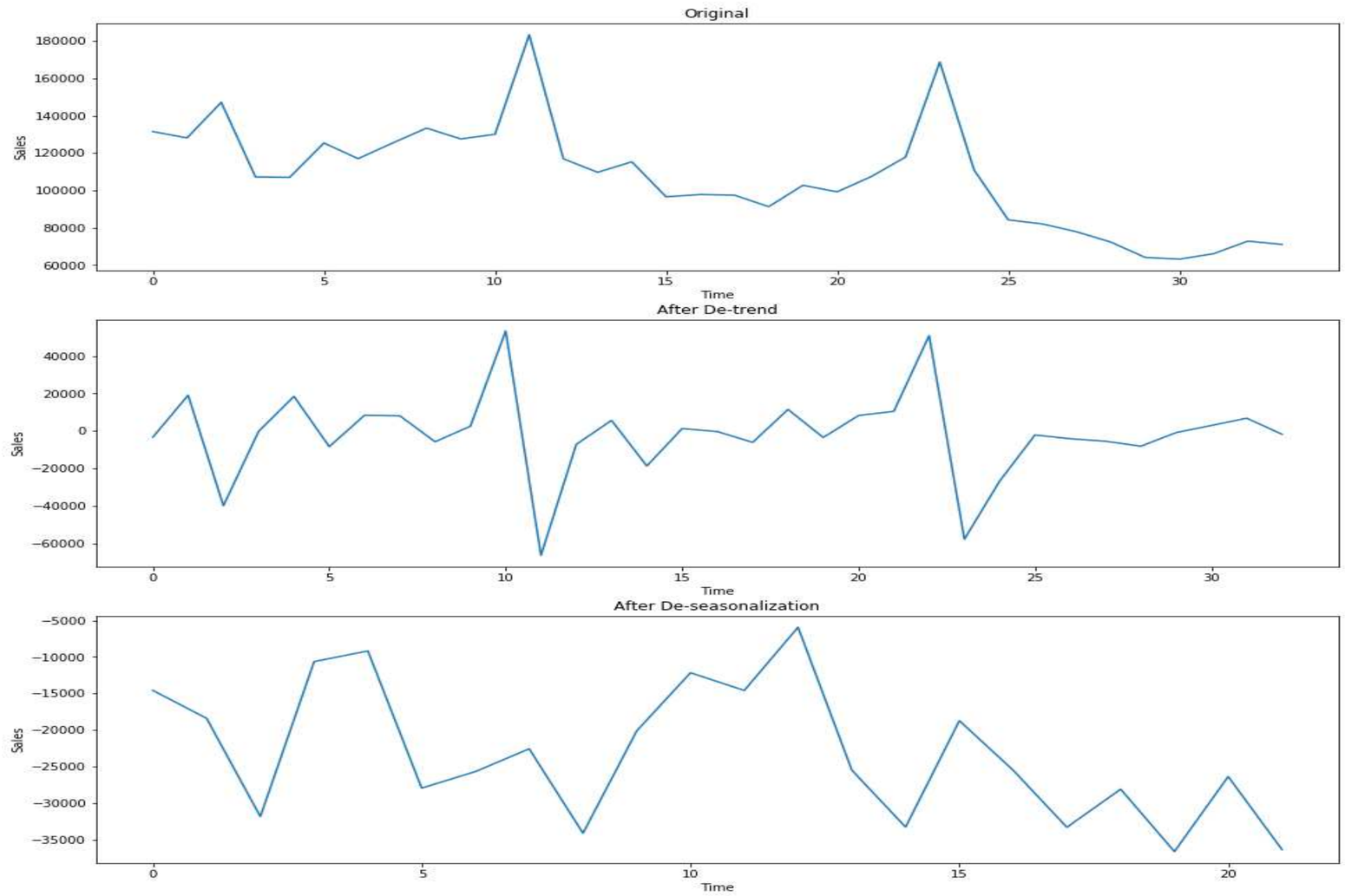


Figure 6: Remove seasonality and trends

Summary

- Now let's check the new P-value.

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.270101
p-value                          0.016269
#Lags Used                       0.000000
Number of Observations Used     21.000000
Critical Value (1%)             -3.788386
Critical Value (5%)             -3.013098
Critical Value (10%)            -2.646397
dtype: float64
```

Figure 7: new stationarity test

- After the transformations, our p-value for the DF test is well within 0.05. Hence we can assume Stationarity of the series.

Conclusion

Summary

- From the above result presentation, we can find that
  There are seasonality and trend in data.

- From the Stationarity test, we can find that
  After removing seasonality and trends, the time series becomes smooth.
  So we can use traditional time series prediction methods for prediction.

Future research

- Predict by traditional time series prediction models such as AR, MA and ARMA.
- Using more models to predict, such as random forests and neural networks.
- Find the most effective model and get my own kaggle ranking.

Thank you & Question