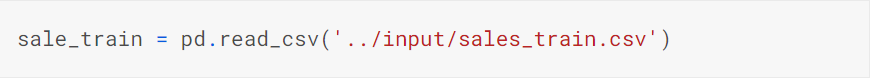
**FUTURE SALES PREDICTION**

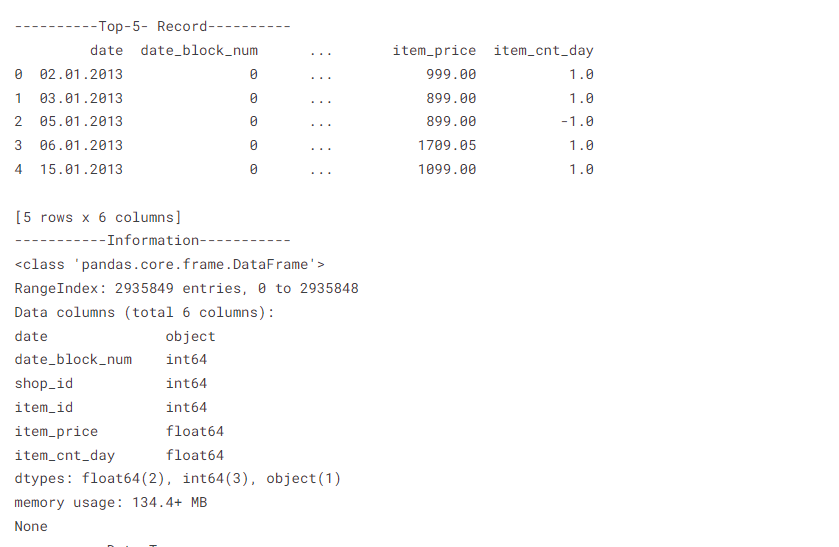
**INTRODUCTION:**

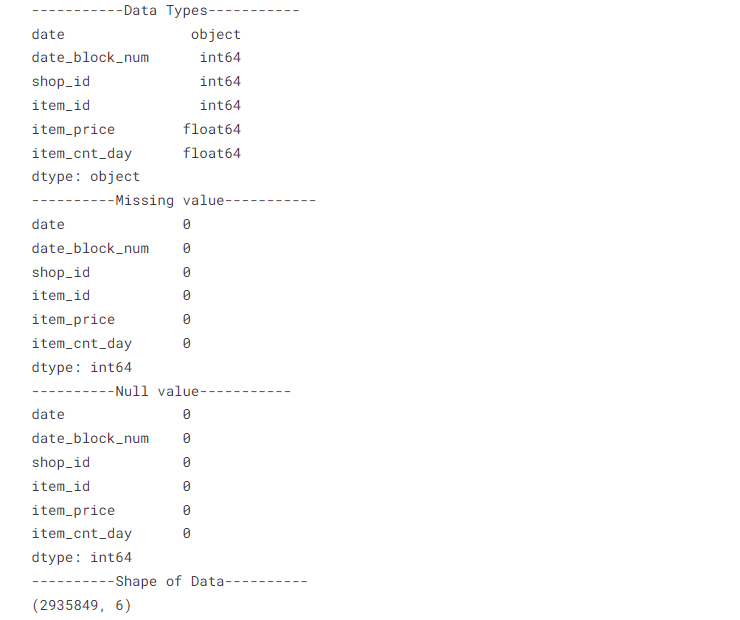
This phase aims to clean, transform, and engineer features in a way that maximizes the model's ability to capture patterns and make accurate predictions. Through careful data preparation and feature engineering, we enhance the quality of input fed into the models. Effective preprocessing lays the foundation for improved predictive models. The given dataset has been pre-processed and the outputs are attached with snap shots.

**LOAD TRAIN DATA**







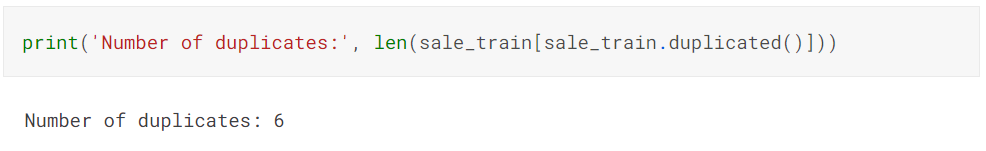


**REMOVING DUPLICATES**

We have duplicated rows, but I don't think that it is a mistake.

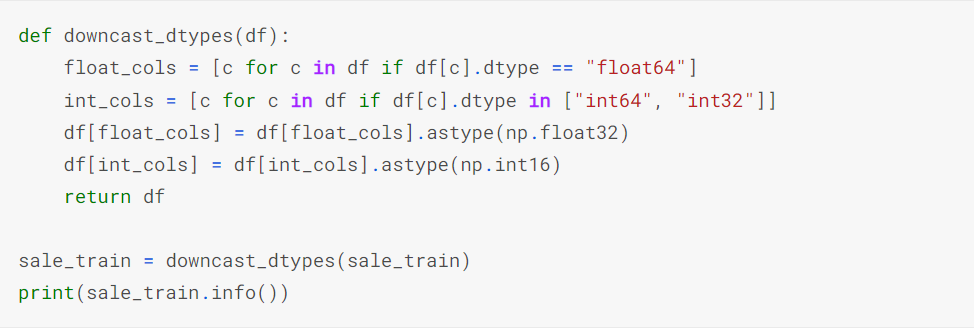
It could be different sales methods or client type, etc.

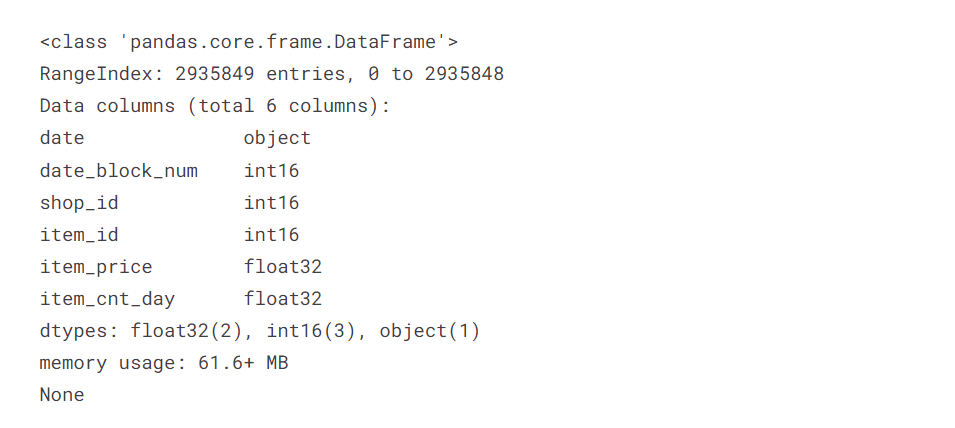
You can remove it, but I really don't believe that 6 rows of 3m can make the difference.



I can advise downcasting our DataFrame. It will save our some memory.

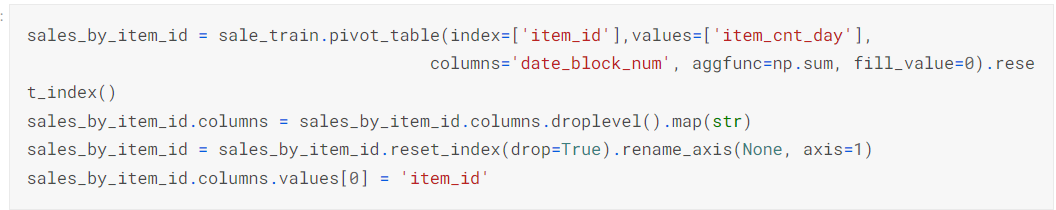
In our case from 134.4+ MB, we went to 61.6+ MB.



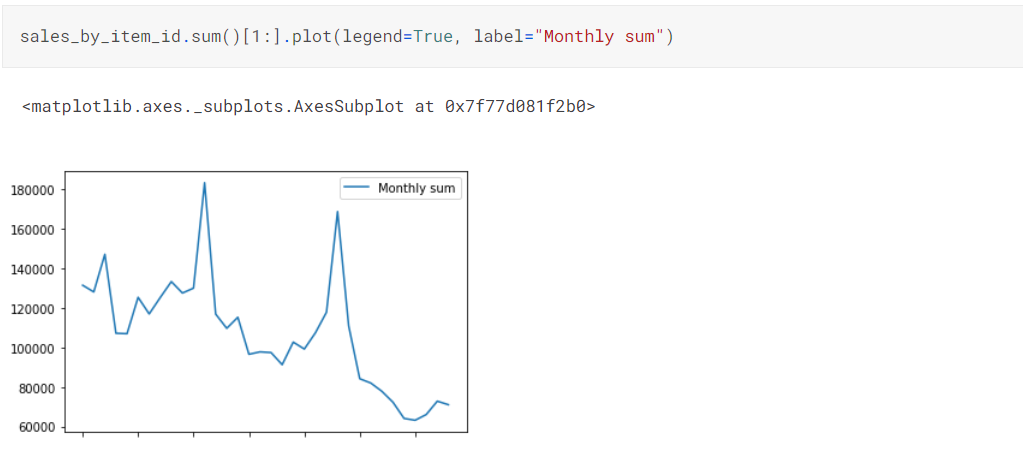


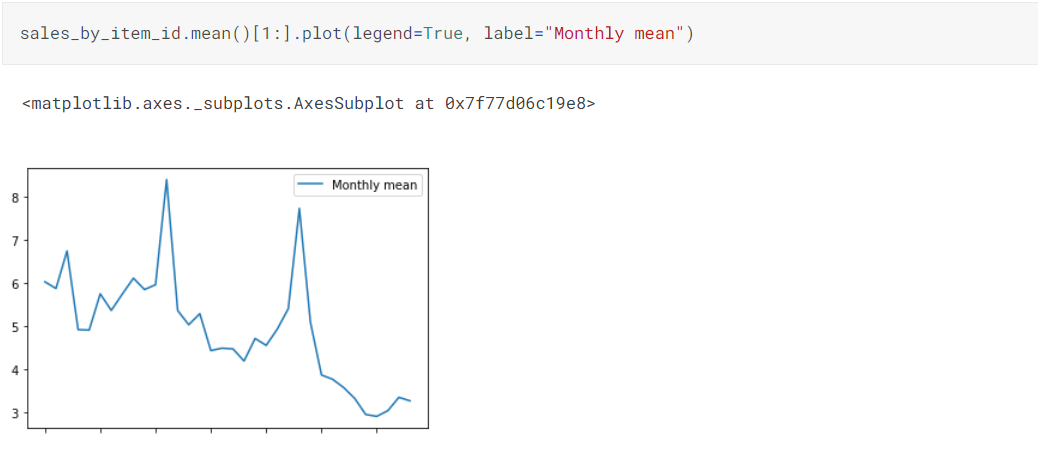
**ITEM\_ID**

Lets group data by item\_id and date\_block\_num and look closer on it.



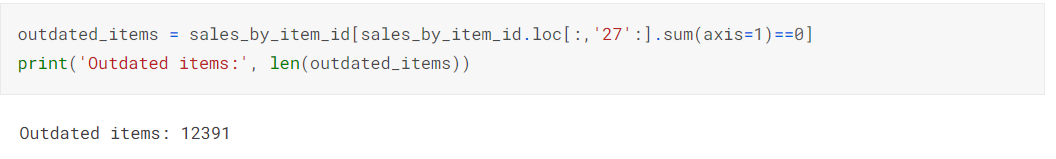
**SIMPLE GRAPH**





### LET'S SEE HOW MANY PRODUCTS ARE OUTDATED (NO SALES FOR THE LAST 6 MONTHS)

12391 of 21807 is a huge number. Probably we can set 0 for all that items and do not make any model prediction.



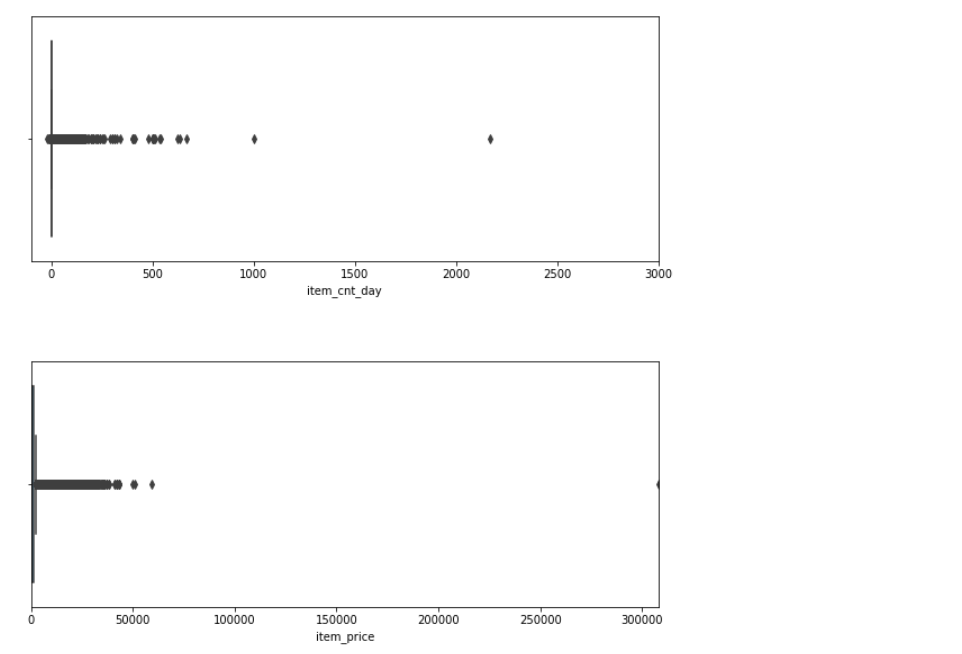
### HOW MANY OUTDATED ITEMS IN TEST SET?

6888 - not much but we have such items



### OUTLIERS BY PRICE AND SALES VOLUME

### 

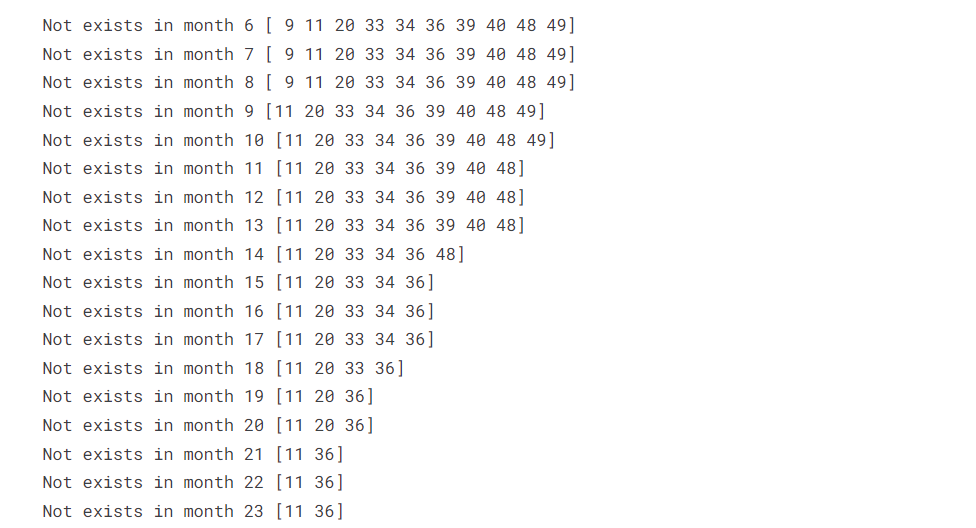


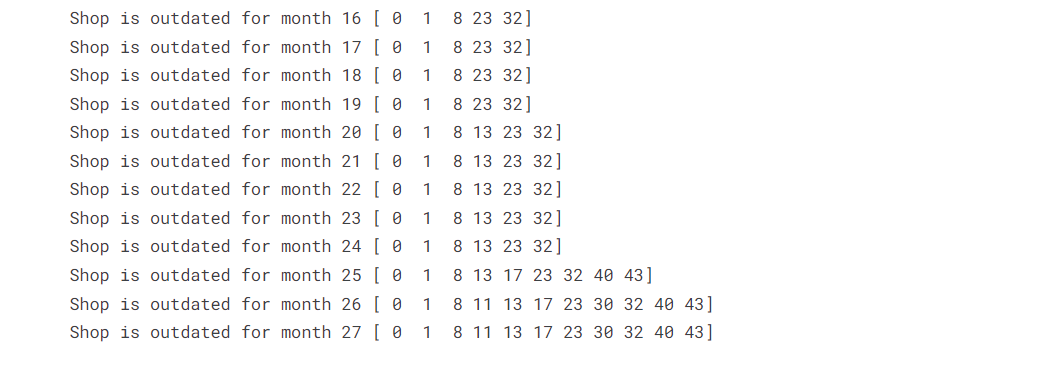
**POSSIBLE ITEM\_ID FEATURES:**

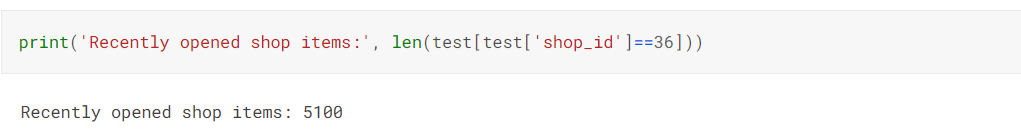
1. Lags
2. Release date
3. Last month sale
4. Days on sale
5. Neighbors (items with id 1000 and 1001 could be somehow similar - genre, type, release date)

## SHOP\_ID

## 







**POSSIBLE SHOP\_ID FEATURES**

1. Lags (shop\_id/shp\_cnt\_mth)
2. Opening month (possible opening sales)
3. Closed Month (possible stock elimination)

**PRICE**

**POSSIBLE PRICE FEATURES:**

1. Price category (1/10/10/20$/ etc.) - obviously (or not obviously), items with smaller price have greater volumes
2. Discount and Discount duration
3. Price lag (shows discount)
4. Price correction (rubl/usd pair)
5. Shop Revenue

**DATES**

**POSSIBLE DATE FEATURES:**

1. Weekends and holidays sales (to correct monthly sales)
2. Number of days in the month (to correct monthly sales)
3. Month number (for seasonal items)

**TEST SET**

The key to my success was the analysis of Test test data.

We have three groups of items:

1. Item/shop pairs that are in train
2. Items without any data
3. Items that are in train



**CONCLUSION:**

In the third phase, the dataset has been preprocessed, which is fundamental to building accurate and reliable predictive models. This involved handling missing values, scaling, encoding categorical features, and possibly applying other transformations like feature engineering or selection. The preprocessed dataset is now ready for the subsequent phases, where it will be utilized to train and validate models