Predict Future Sales

Final Report

This document describes the various methods applied to daily historical sales data and its prediction.

S Shankar

Problem Statement/Objective:

Predicting future sales(forecasting) helps the company to gauge its revenues for the immediate future. This can be done monthly, quarterly or yearly, depending on the definition of the sales period. It helps the company to manage the supply chain, cash flow and perform strategic planning.

- Inaccurate sales forecast will prove costly to the company with inventory purchases, which is tied to forecasted sales. Low inventory levels will result in placing rush order to the vendor. Similarly, over stocking will also cost the company as cash sits idle and held up in the inventory.
- An accurate sales forecast helps the company to achieve faster revenue growth and higher margins. Accurately forecasting the sales and building a sales plan can help avoid unforeseen cash flow problems and manage the production, staff and financing needs more effectively.

A sales forecast is an essential tool for managing a business of any size.

Our objective is to work with a challenging time-series dataset consisting of daily sales data, kindly provided by one of the largest Russian software firms - 1C Company.

We have to predict total sales for every product and store in the next month.

Potential Clients

The potential clients for this project would be any retail stores, larger corporations and also vendors and dealers of any products.

Exploratory Data Analysis

Missing Data

There are no Missing data - NaNs in our train data set.

Outliers

There are outliers present in the item_cnt_day and item_price of train data set.

Datatype

Date column of train data has been converted to pandas datetime format.

Duplicates

There were 6 duplicates were found and looks like they were bought twice due to their popularity.

Negative item sold count

We see -1 in th count , assuming that the product was returned.

One item was sold as -1, we took the mean price for that item.

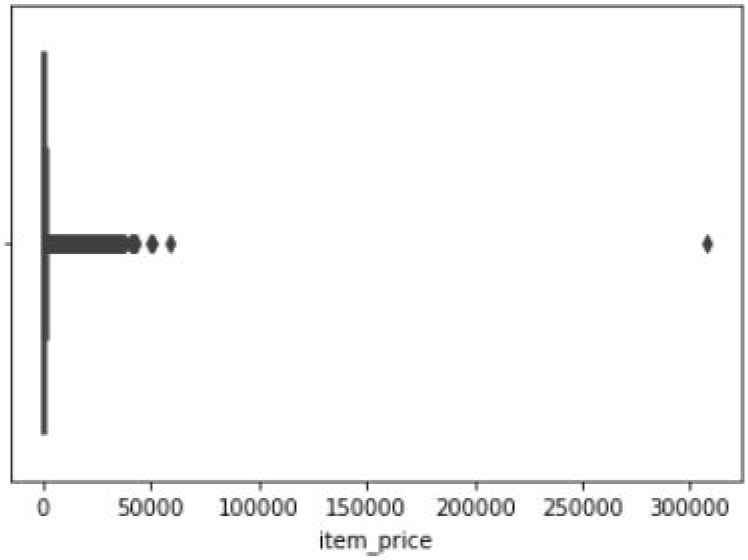
Translation

On translating from Russian to English, we see that

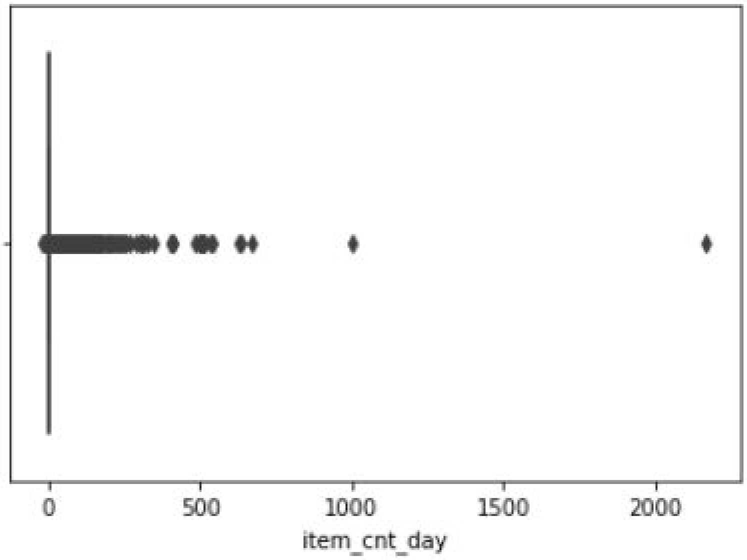
Item Categories contains type and subtype with ‘-’ delimiter

Shops contains city with space delimiter

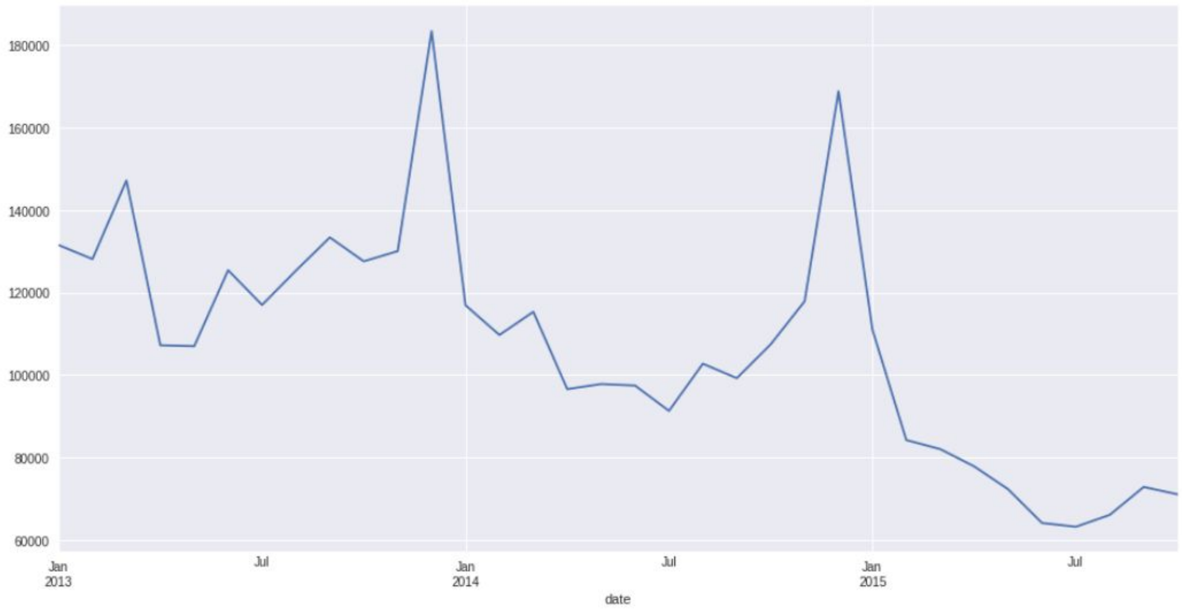
Visuals



This plot shows the item_price outliers [11365, 6066, 13199] present in training data set

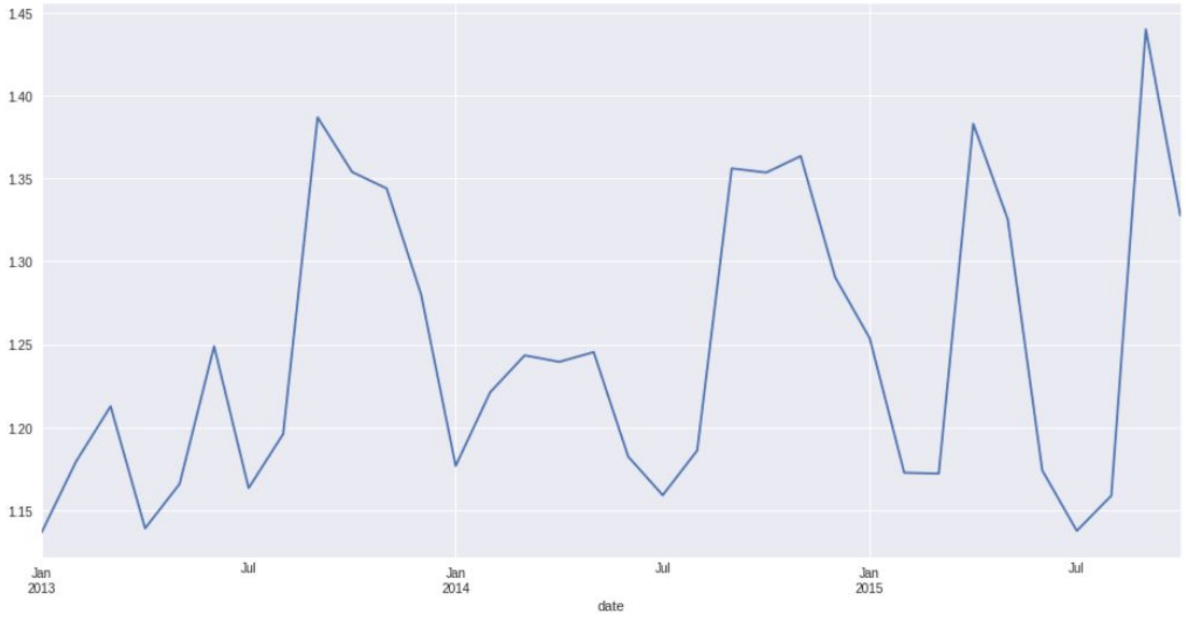


This plot shows the Item count day outliers: [8057, 20949, 9242, 19437, 3731, 11373, 9249, 9248] present in training data set

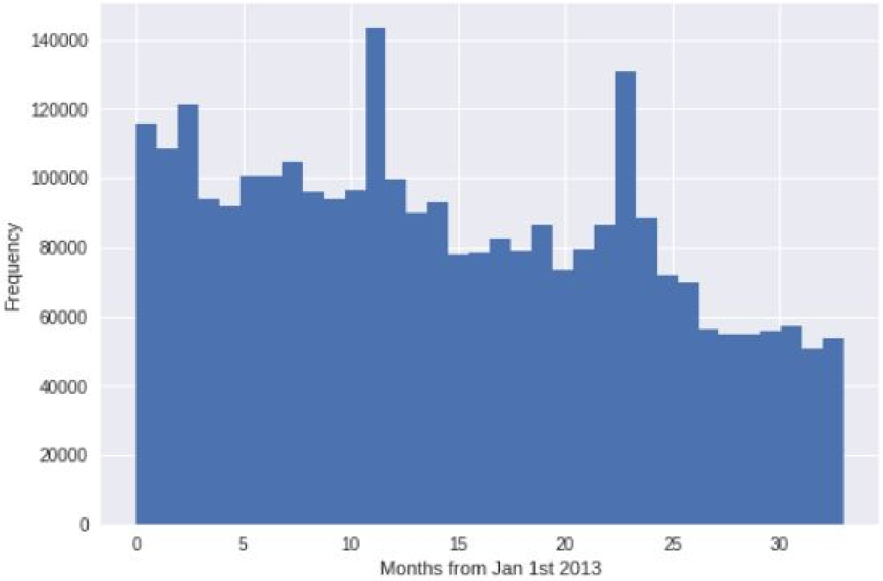


Monthly sampling distribution on the train data(Count of items sold per month)

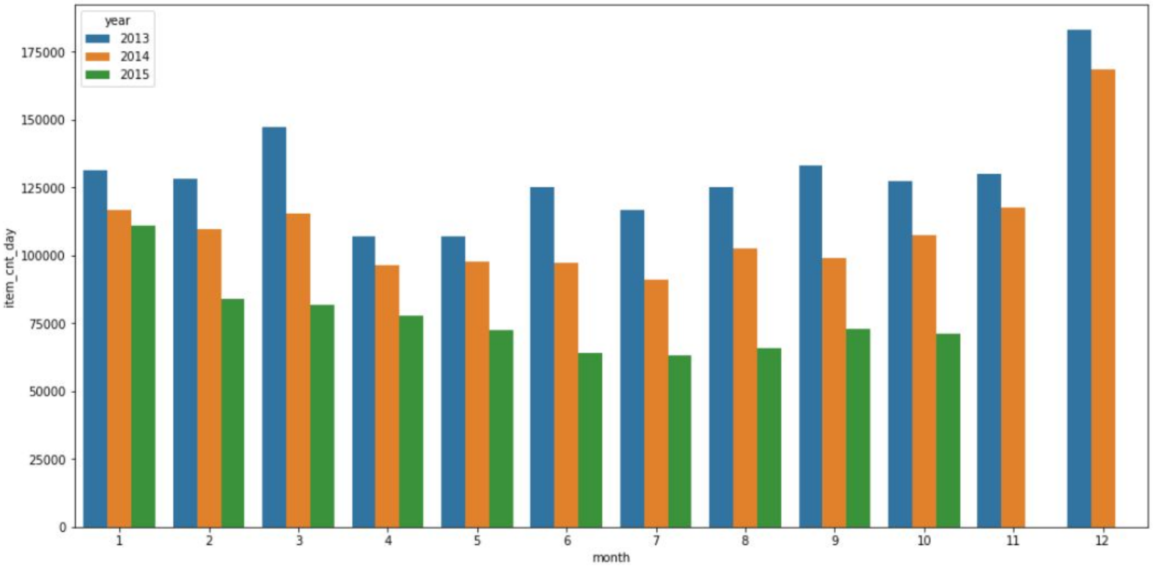
There is a sharp increase in the number of items sold in December(both 2013 and 2014), this is the holiday period. But there is a negative trend, i.e. general decrease in sales (even in the holiday period)



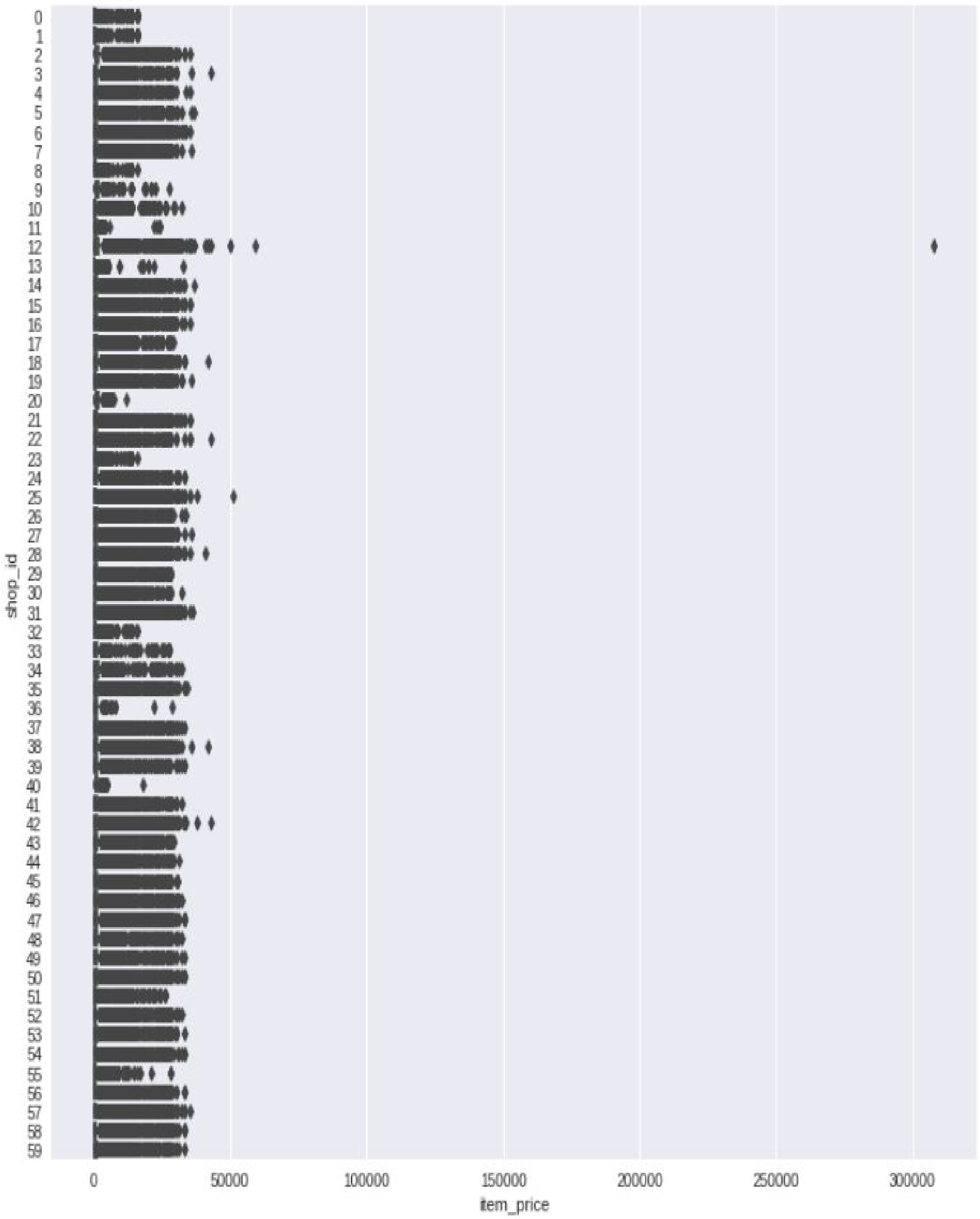
Mean of items count per month



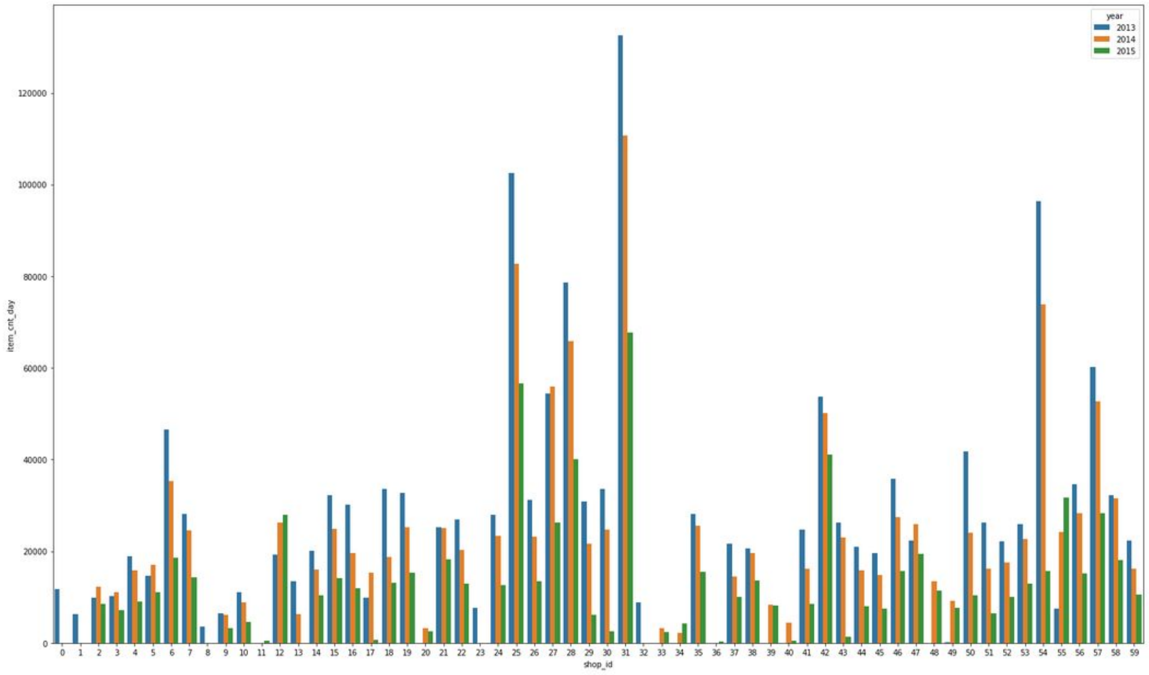
Histogram items sold per month



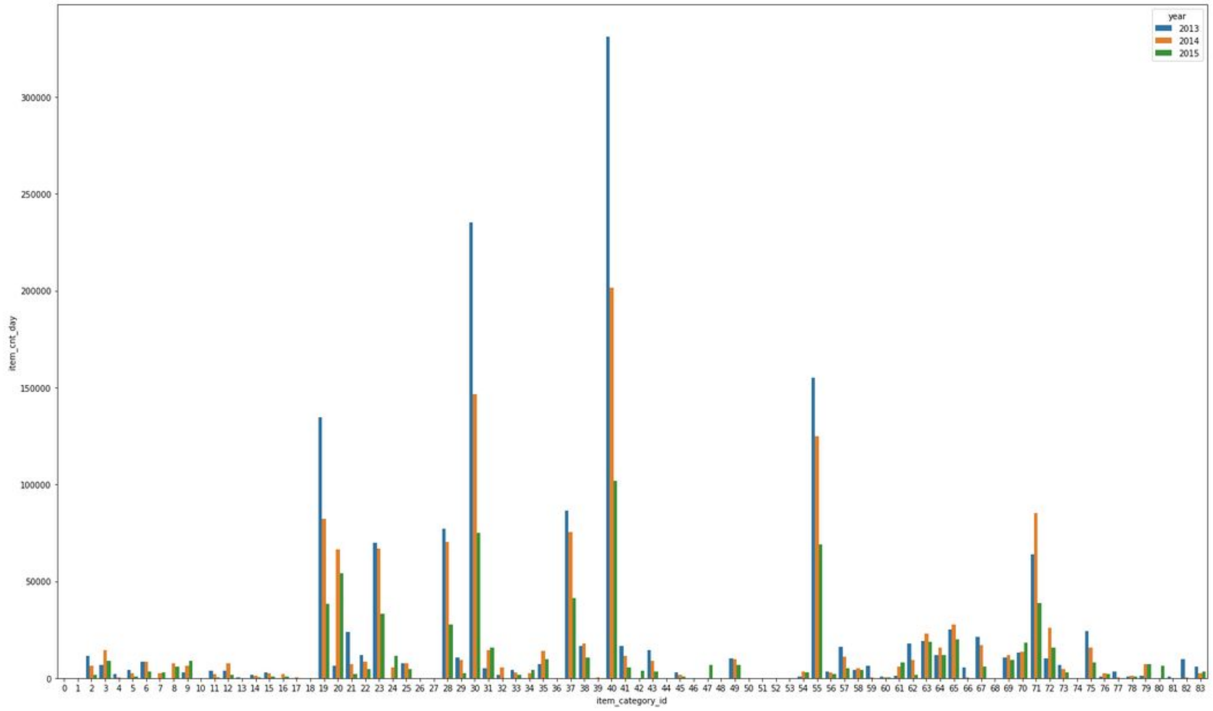
As observed earlier, there is a negative trend on sales(month vs sum of item count)



Item prices per shop (The most expensive item is Radmin 3 - 522 лиц.)



Some shops performed significantly better than others , especially 25,31,54



Some categories clearly sold better than others , like 30,40,55

General Observations

- There are 22170 items in the catalogue and 60 shops
- The total possible number of combinations 22170*60 = 1330200
- The train set consists of 60 unique shops and 21807 items
- The test set consists of 42 unique shops and 5100 items
- There are 84 different categories in the catalogue
- There are 2,14,200 items altogether in the test file, we need to predict items sold per month for the 60 shops
- item_cnt_month from samples_submission gives us the idea of how many items are sold in a shop per month(this is given in test)¶
- The data is for 34 months , for 60 shops and around 2.9 million(2935849) items
- Train data is Daily Historical data from January 2013 to October 2015
- We see that the highest sales occurs in the month of december
- The reason for the decline of sales could be that year on year could be that these items could be electronic or software products , where a newer version was released and hence the older versions saw a decline
- The Item catalogue could also not have been updated, where as the items were removed from the shops and not available for sale
- Some shops perform better than the rest combined. Could be its location in a High street or could the shops would have started or shut down(closed) in the given period
- Some item categories were clearly popular in sales

Inferential Statistics:

Dependent Variables

Sales data, which is Item unit times prices

Independent Variables

Time period and the dummy variables(on item categories, shops)

The item categories and shops dataset contains categorical data.

The item categories dataset contains 84 unique item category values. They can be further interpreted as category type with 22 and category subtype with 66 unique values.

The shops dataset contains 57 unique shop values. They can be further interpreted with location as city with 31 unique values.

Statistical Tests

In the training data set the average sales price was: 932.1405, total items sold: 3646036 and total sales value: 3398617900.

Single sample t-test: Average price per sale is equal to 800

- Null Hypothesis H0 - Average price per sale is equal to 800
- Alternative Hypothesis H1 - Average price per sale is not equal to 800

Interpretation: Since p-value is 0, H0 is rejected. We conclude that Average price per sale is not equal to 800.

Independent sample t-test: Average price per sale of the cities Москва & Н.Новгород are compared

- Null Hypothesis H0 - Average price per sale is the same for the two cities - Москва & Н.Новгород
- Alternative Hypothesis H1 - Average price per sale is not the same for the two cities - Москва & Н.Новгород

Interpretation: Since p-value is nearly 0, H0 is rejected. We conclude that Average price per sale of the two cities are different.

ANOVA: Average sales is different for 12 months

- Null Hypothesis H0 - Average sales for 12 months are same
- Alternative Hypothesis H1 - Average sales is different for atleast 1 month

Interpretation: Since p-value is 0, H0 is rejected. We conclude that Average sales is different for at least 1 month.

Chi square: Number of total Item sales per month is Uniformly distributed or not

- Null Hypothesis H0 - Number of total Item sales per month is Uniformly distributed
- Alternative Hypothesis H1 - Number of total Item sales per month is not Uniformly distributed

Interpretation: Since p-value is 0, H0 is rejected. We conclude that Number of total Item sales per month is not Uniformly distributed

Linear Regression:

Predicting Average sales price per sales from month(date_block_num).

The estimated regression line is

avg_sales_per_block = 809.3849 + (22.6911 * date_block_num)

The regression model is significant , F = 28.07, P = 0.000

Regression equation is used to predict the next two month blocks - 34 & 35

Predicted avg_sales_per_block for the block 34 is 1580.8823 and 35 is 1603.5734

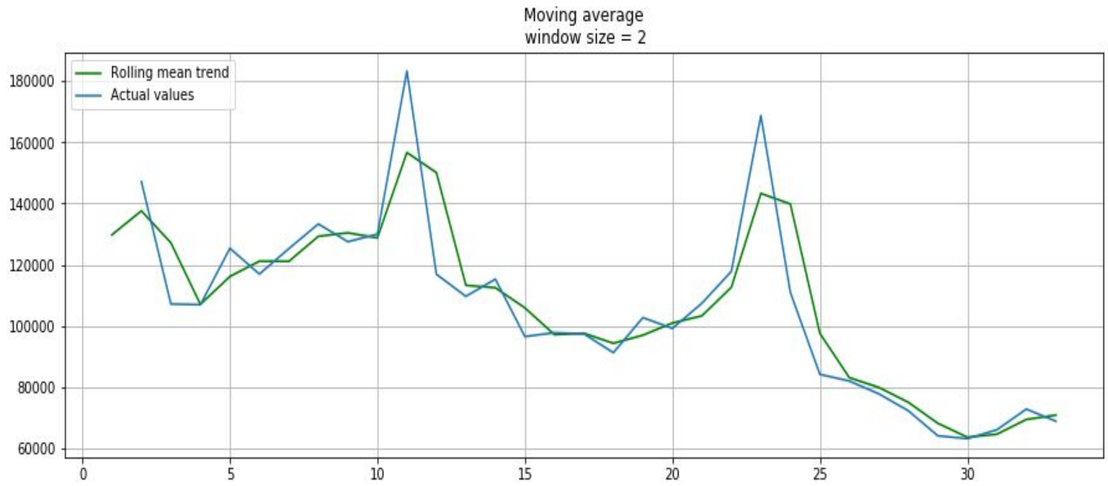
Building the Model:

Linear Regression:

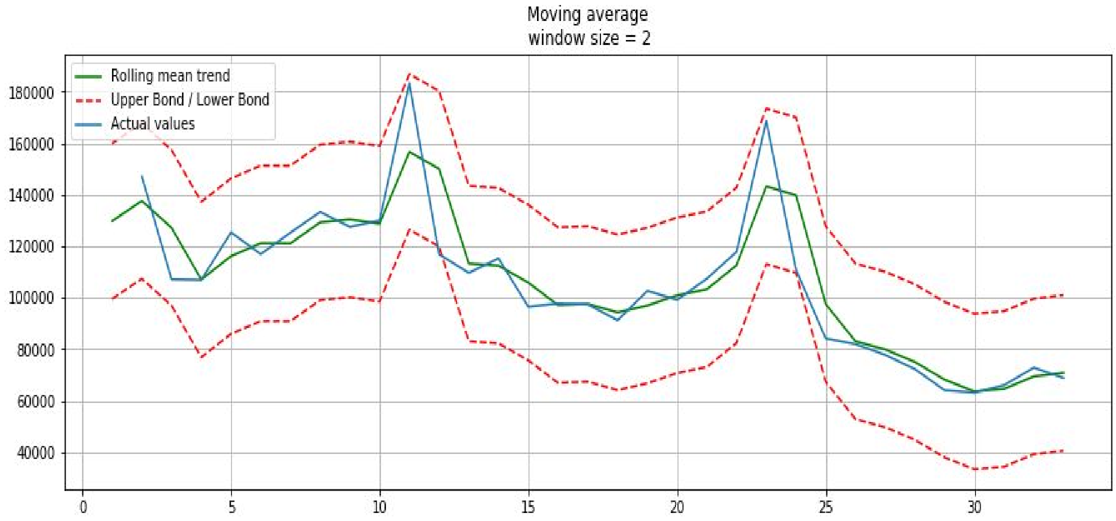
Built a simple model and estimated regression line is avg_sales_per_block = 809.3849 + (22.6911 * date_block_num) The regression model is significant , F = 28.07, P = 0.000 Regression equation is used to predict the next two month blocks - 34 & 35 Predicted avg_sales_per_block for the block 34 is 1580.8823 and 35 is 1603.5734

Move, Smooth, Evaluate in a Time series:

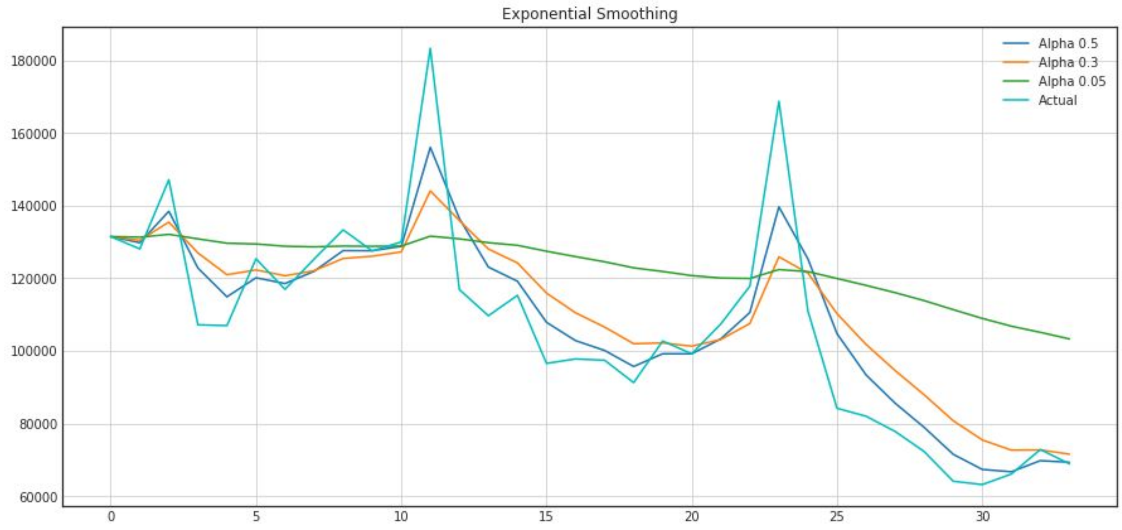
Moving average Smooth by previous 2 months



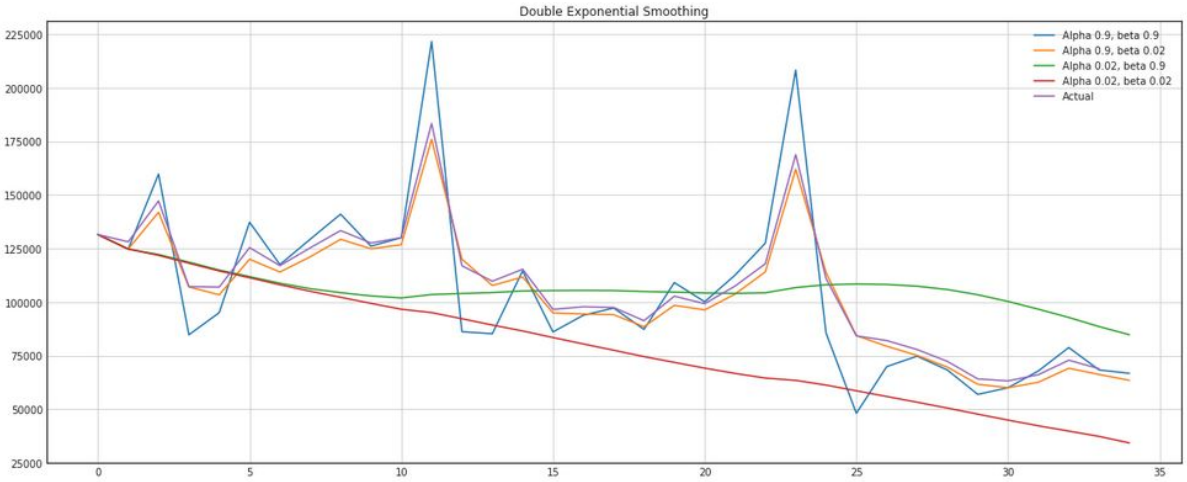
Moving average with confidence intervals



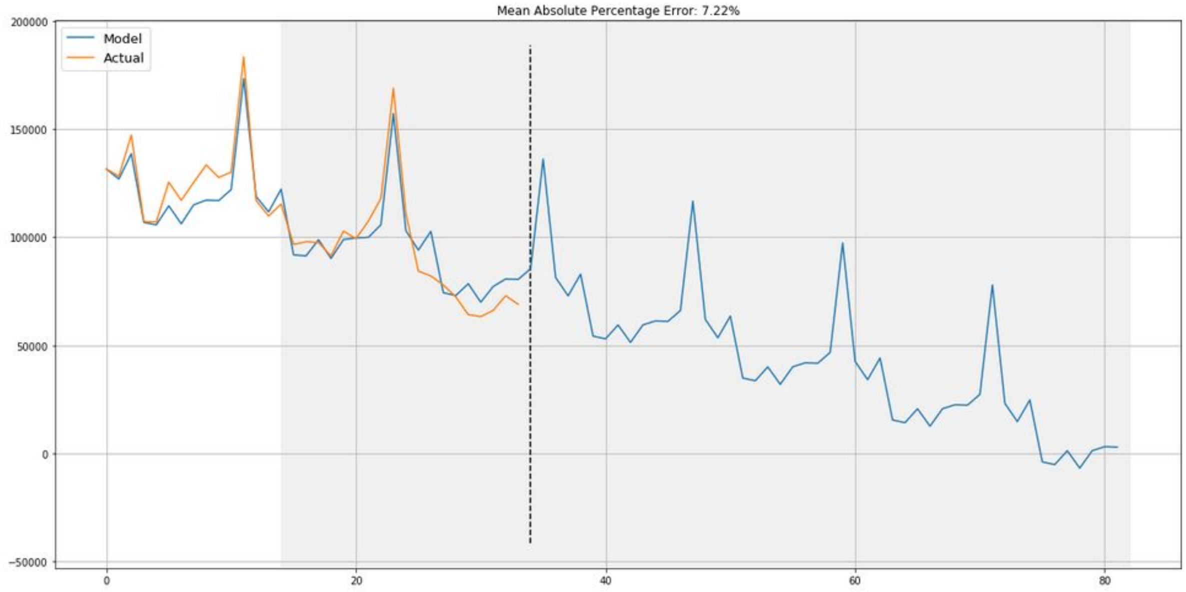
Exponential smoothing of the moving average

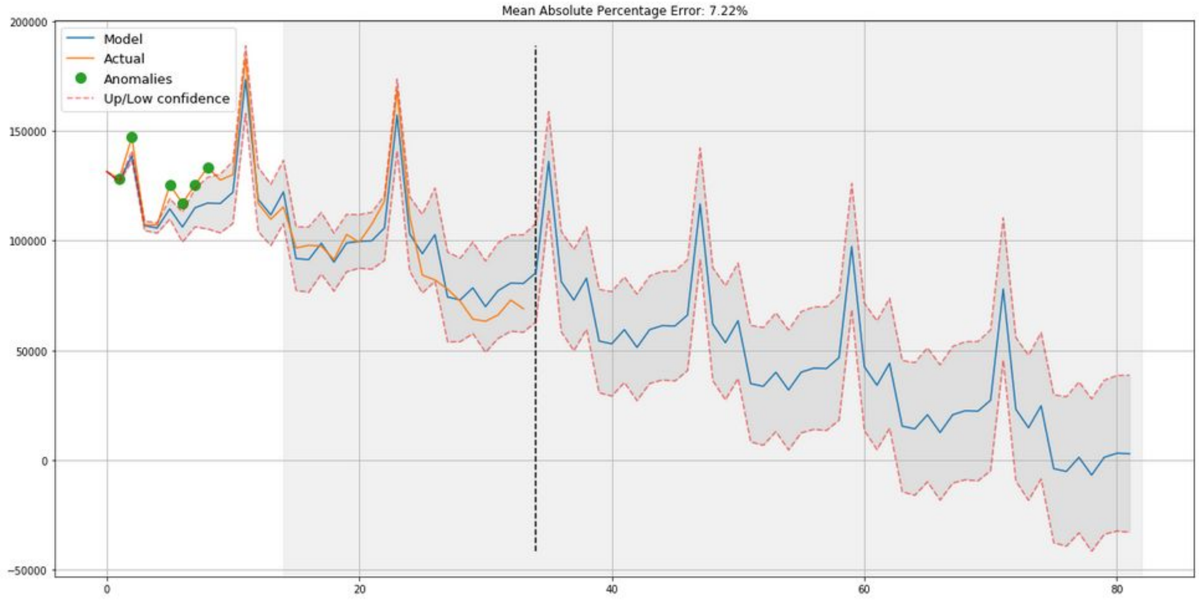


Double Exponential smoothing of the moving average



Triple Exponential Smoothing also known as the Holt-Winters method

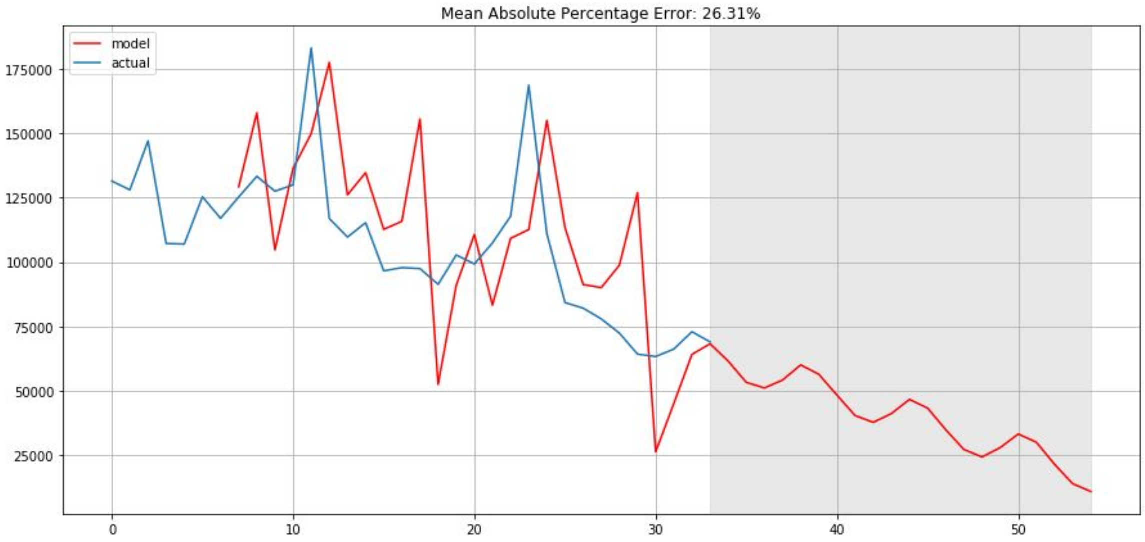




Judging by the plots, our model was able to successfully approximate the initial time series, capturing the daily seasonality, overall downwards trend, and even some anomalies. If we look at the model deviations, we can clearly see that the model reacts quite sharply to changes in the structure of the series but then quickly returns the deviation to the normal values, essentially "forgetting" the past. This feature of the model allows us to quickly build anomaly detection systems, even for noisy series data, without spending too much time on preparing the data and training the model.

SARIMA model (making the series Stationarity)

SARIMA: Seasonal Autoregression Moving Average model



We got inadequate predictions. Our model was wrong by 26.31% on average, which may not be appropriate to predict.

Boosting

- XGBoost

  The submission on Kaggle scored 0.90213 in the public leaderboard

- CatBoost

  The submission on Kaggle scored 1.04083 in the public leaderboard

Future Scope

- Attempt different methods to clean the data.
- Attempt different methods to explore and visualize the data.
- Explore different regression models or combinations of such models.
- Limitations in memory meant that larger dataset with multiple features could not be modelled. Try sequential or other methodologies available to run the larger datasets like Dask, Wowpal wabbit.
- Attempt with different models ensembling/stacking.