Report of the project

1.Idea

We are given data from Walmart supermarket. It consists 3 files:

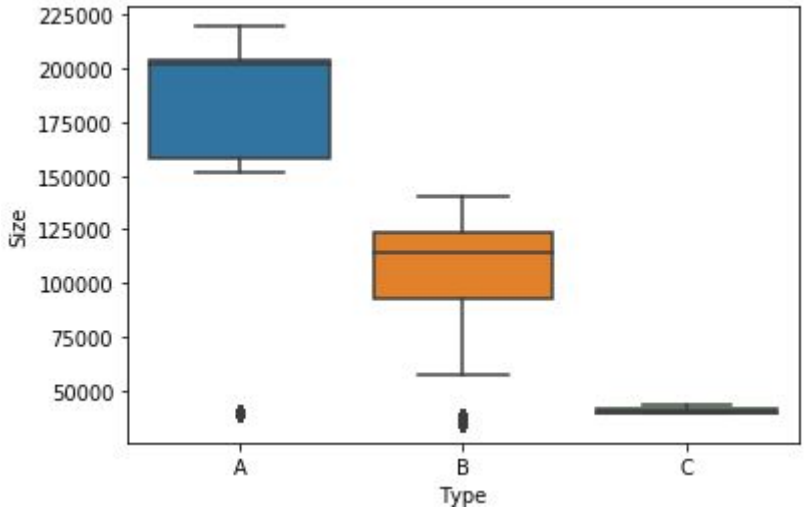
- file features additional data related to the store, department, and regional activity for the given dates.
-file containing anonymized information about the 45 stores, indicating the type and size of store.
-file containing historical training data, which covers to 2010-02-05 to 2012-11-01. It is our training data. It contains information about date, store, weekly sales and information if in that day was holiday.
- file containing the same information as in training set without weekly sales and date from 2012-11-02 to 2013-07-26.

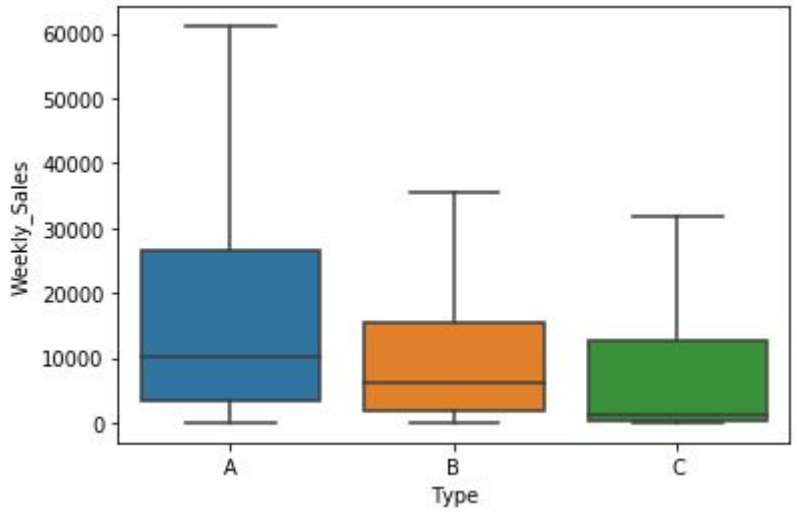
We want to predict weekly sales in given shops for given dates. It is a question for Kaggle competition and I will check results of test set using their score.

2.Approach

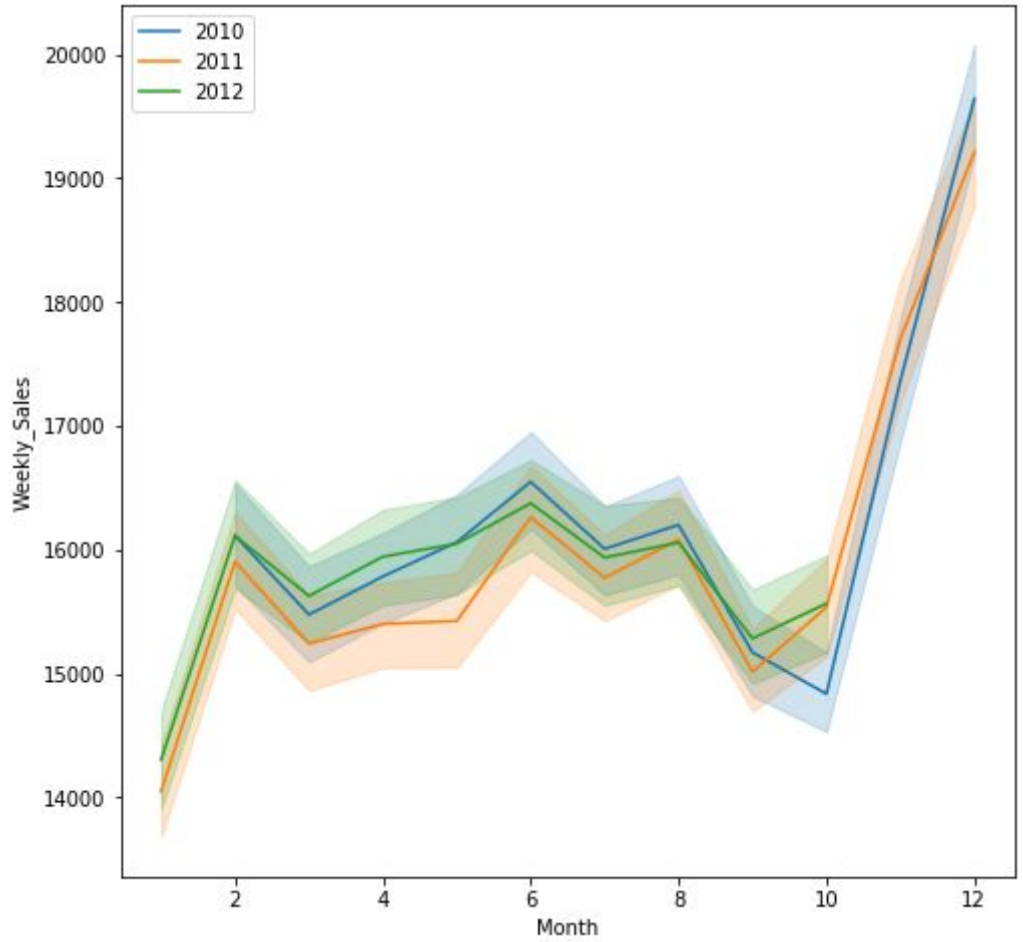
At the beginning I want to conduct Exploratory Data Analysis. At the beginning I checked distribution of every feature and found out few interesting and important conclusions:

- The bigger shops are, the higher weekly sales they conducted.

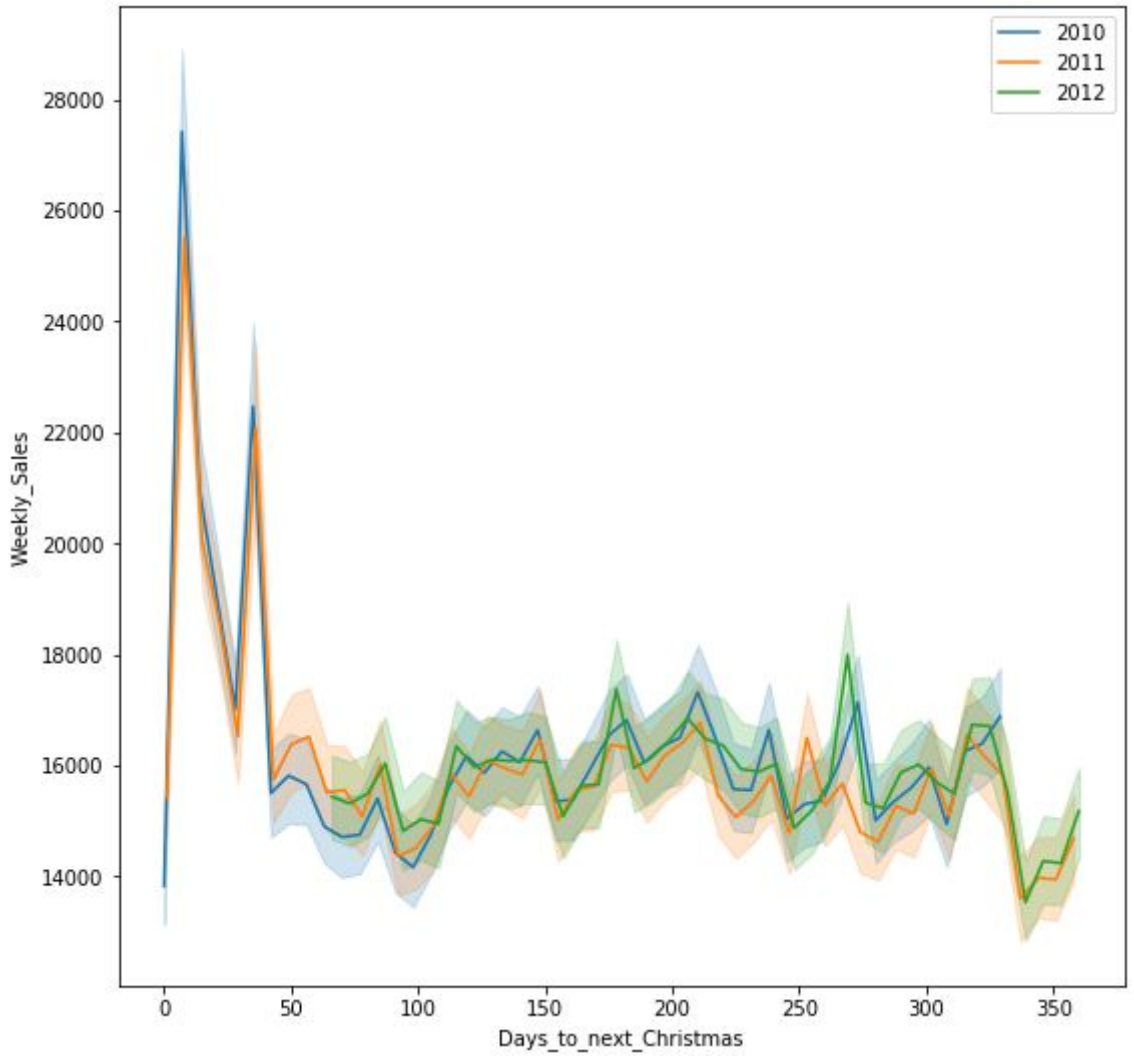




- We see some trends that CPI get higher during year and Unemployment lower but it is not as essential as other.
- One of the most important thing is the fact that weekly sales are nearly the same at each date on every year, so it shows us that data is seasonal.



- Last important conclusion after EDA is the impact of few holidays, especially Christmas in weekly sales, so it could be a good idea to have information of days to Christmas as new column.



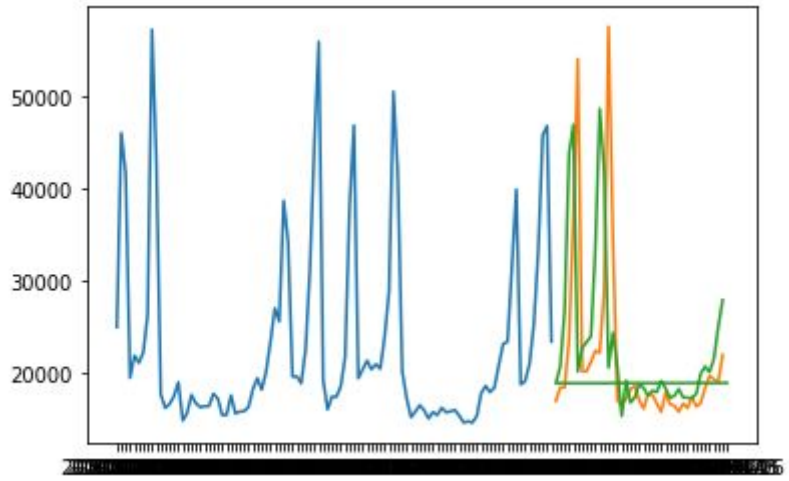
Checking some basics models

I tried to compare 3 different models: Linear Regression, Random Forest and XGBoost. I created a test set from train set and checked performance. As Linear Regression had a very big mean absolute error, I concluded that it is not great model to use it because it didn't catch seasonality of data. Random forest and XGBoost had a great result on our test set, but it worked poorly on completely new data in original test set. It is because those two models are good in interpolation where we know previous and following data so they can predict in a nice way the "gaps" in it. However the Kaggle score was very bad! MAE of prediction was more than 23000! That prove my theory about ability to forecasting by these models. We make it better by predicting values of

features in feature and shop files but it will not help these models to make as much impovement as we want.

Optimal algorithm

Optimal algorithm is called **Exponential Smoothing** and it is simply algorithm where we detect trends and seasonality of our data. Unfortunately it only take under consideration our Date and Weekly Sales so we don't use another features which may be helpful. Nevertheless it has quite good results. This diagram shows how it work:



Where green line is our prediction and yellow line is our test set. It turns out that we have nearly 5 times lower MAE so could be happy with the results!

Next steps:

We could modify exponential smoothing to take under consideration other features, especially "Days to Christmas". It will surely make our error even less.