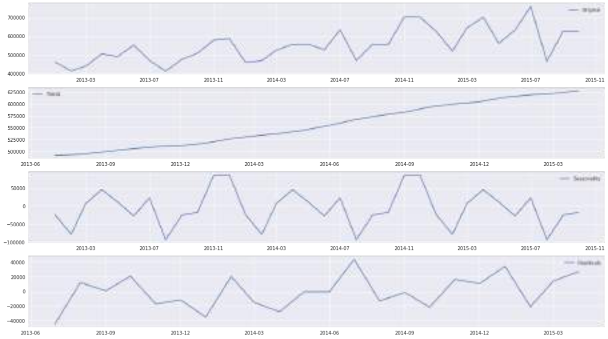
Approach

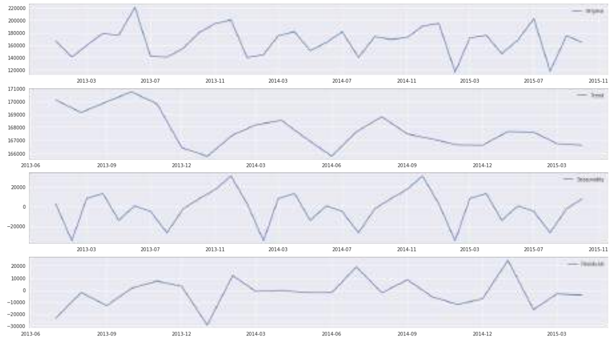
Quality checks performed / Errors found:

- No Error were found.

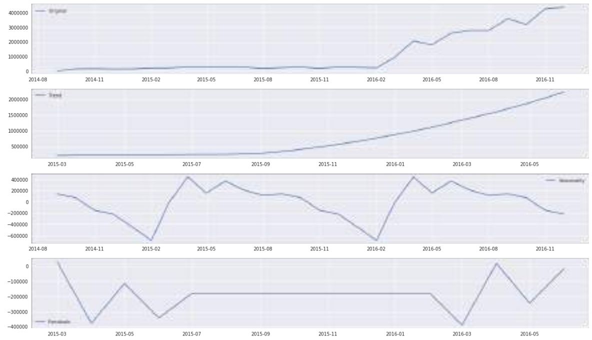
Key observations / Trends:



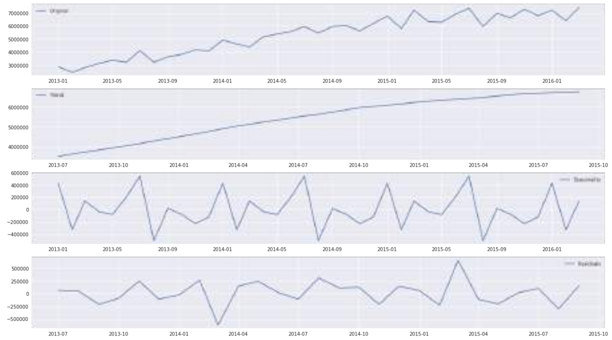
England Product - 5 Sales



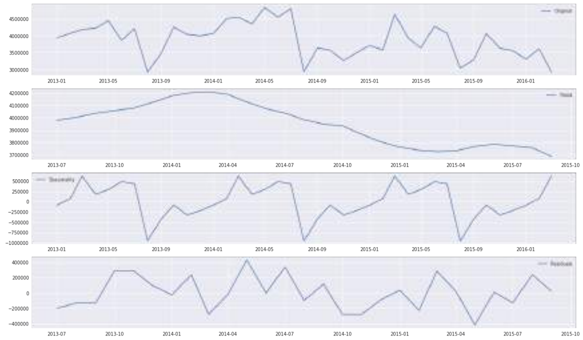
England Product - 4 Sales



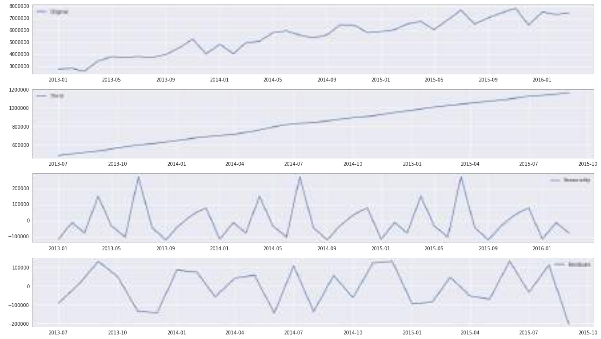
Columbia Product - 3 Sales



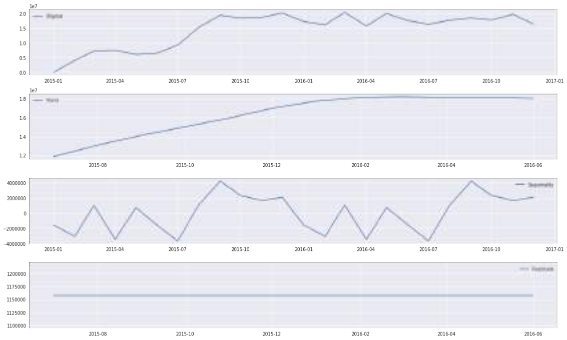
Columbia Product - 2 Sales



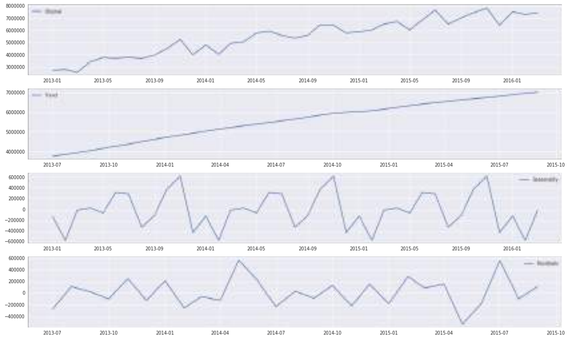
Columbia Product - 1 Sales



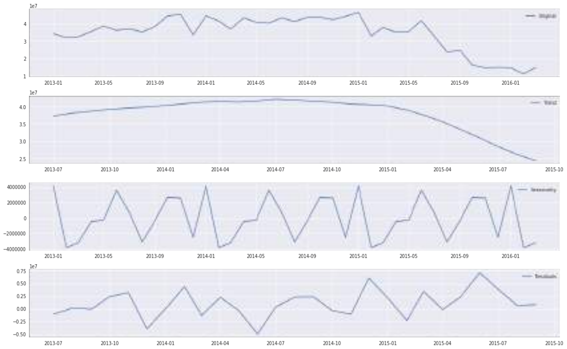
Belgium Product - 2 Sales



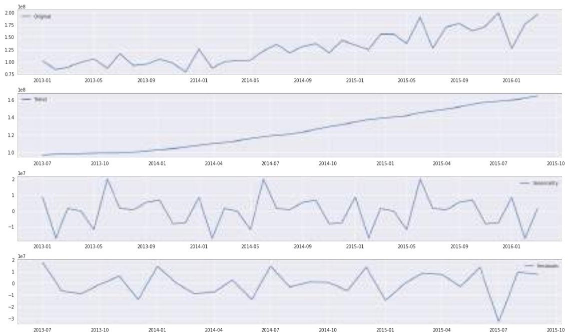
Argentina Product - 3 Sales



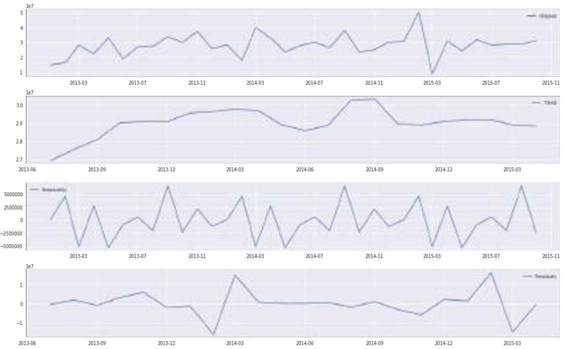
Argentina Product - 2 Sales



Argentina Product - 1 Sales



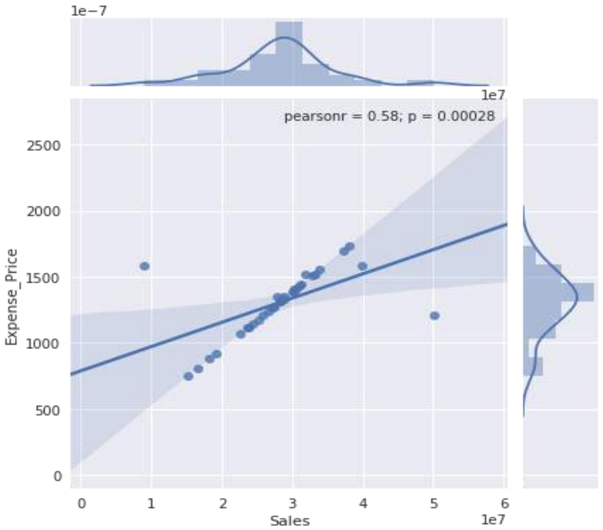
Denmark Product - 2 Sales



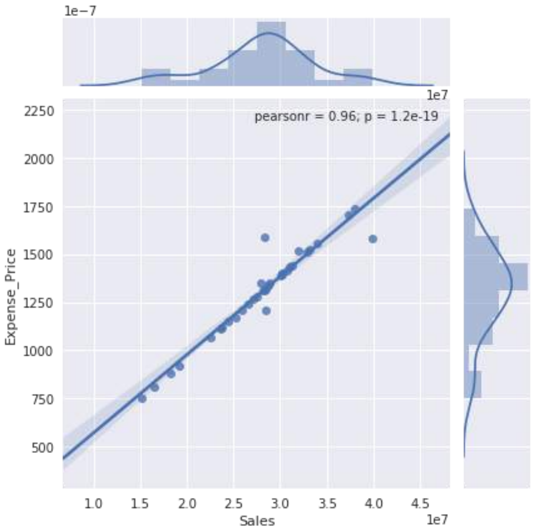
Finland Product - 4 Sales

CORRELATION TREND:

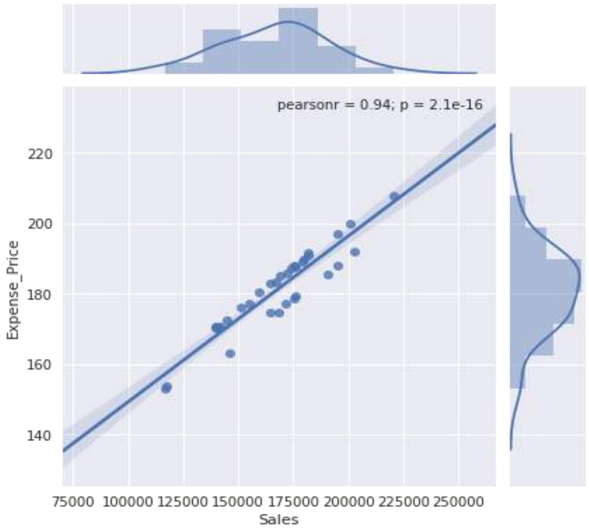
EXPENSE PRICE DATA VS SALES DATA:



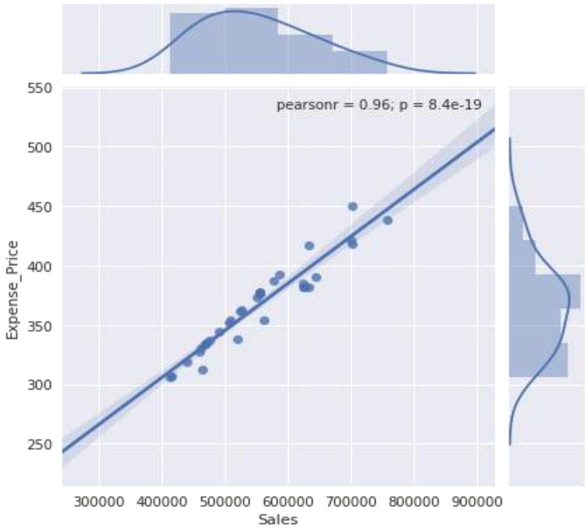
Finland Product - 4 with Outlier



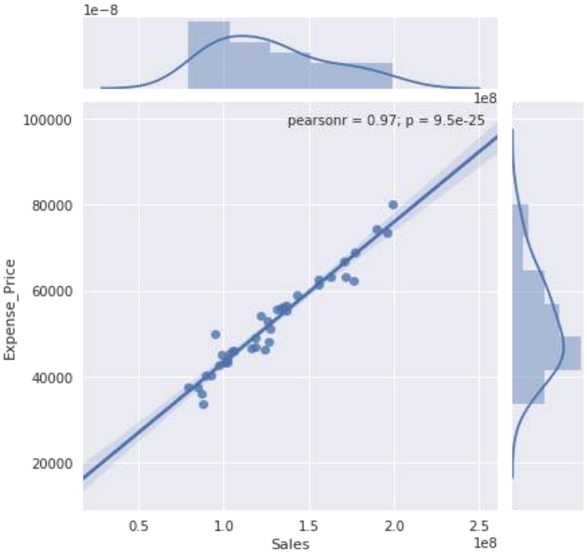
Finland Product - 4 without Outlier



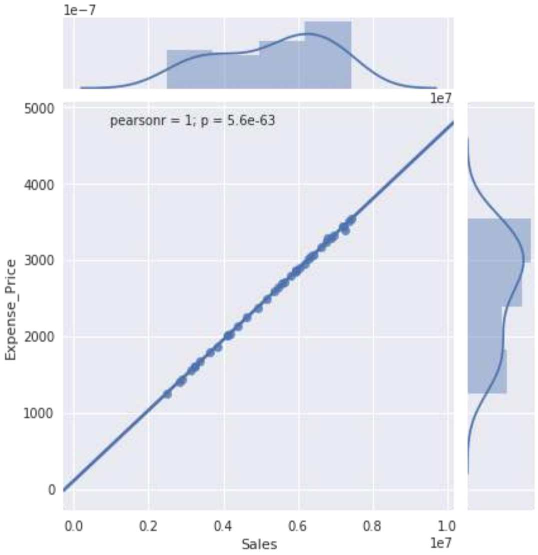
England Product - 4



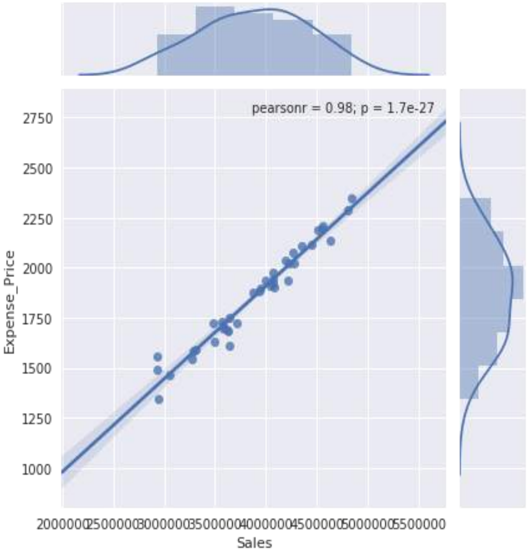
England Product - 5



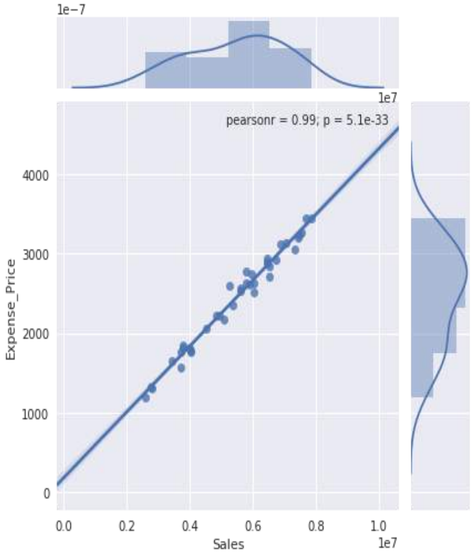
Denmark Product - 2



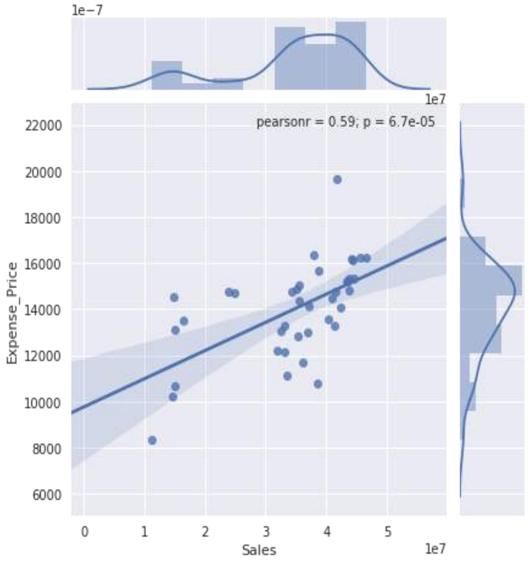
Columbia Product - 2

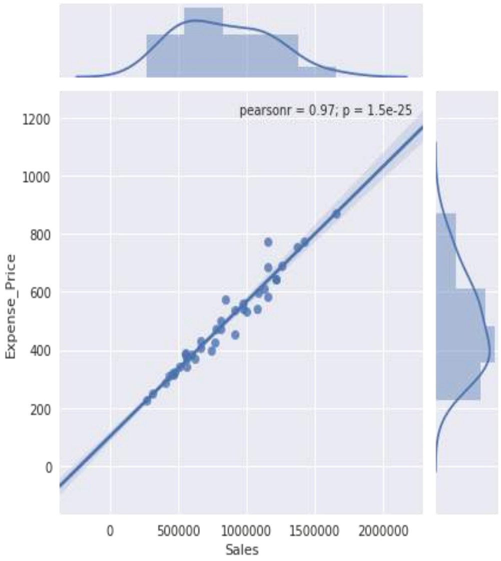


Columbia Product - 1



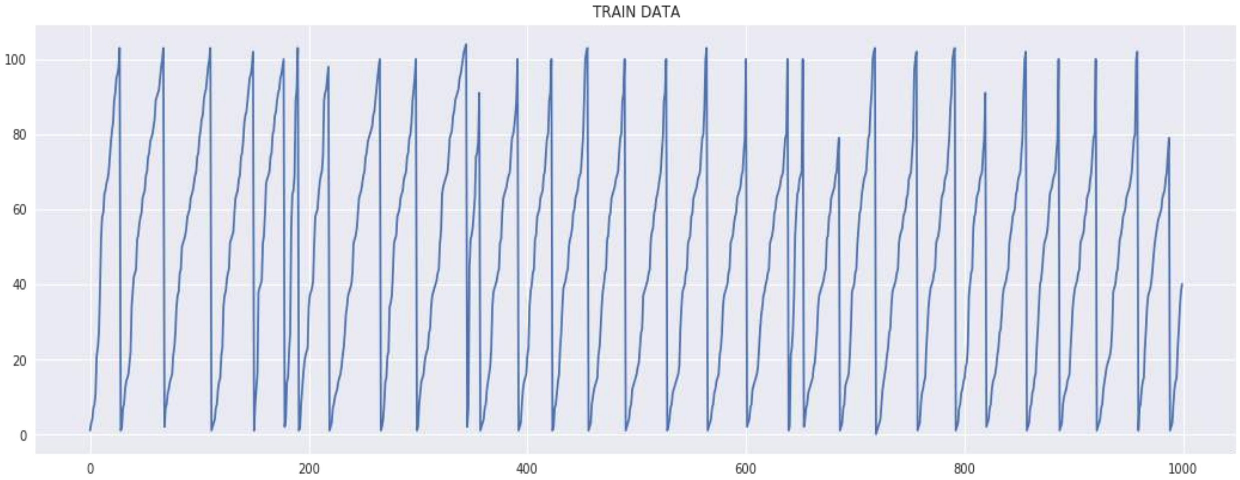
Argentina Product - 2





Argentina Product - 1

Belgium Product - 2



Graph of Merchant ID with Time. There is a trend we can observe that Merchant ID increase linearly and drop and then increase. The Point where the ID drops marks the end of the Week in the dataset.

Data preprocessing steps:

- In the data preprocessing step, what I did is, I just **took sum of all sales record for every particular year and month.**
- The above preprocessing is done to ensure that the sales data matches with the data given in the promotional expense dataset, as because from above we can see that their is a clear cut r**elation between expense price and sales data. They are highly correlated.**
- Then after the above step is done I used year and month columns to create timestamp which will be used as the index for our problem.

Model choice explanation:

- I tried different ML models like Random Forest, Linear Regression, SVM, and Decision Tree but they didn't give me the results I was expecting as some model were underfitting like RF and Linear Regression and some are overfitting like Decision Trees so after all this I tried **XGBOOST** and tuned it with different values for the parameter and it just improved my score with a hope that it would not overfit.
- XGBOOST is used where we have both the expense data and sales data for a particular product.
- I have **used ARIMA where there is no expense data available** to predict the future events with the help of timestamp created in the preprocessing step.
- I have used ARIMA because it can tackle outliers and predict the future events quite decently without overfitting.

Expected error for submission:

- As we have only few limited data points for each product, therefore **chances of overfitting** increases extensively and this may lead to incorrect predictions and thus expected error increases.

Top 5 most significant variable in model:

I only found these 3 features to be contributing the most:

- **Expense Price**
- **Year**
- **Month**