Data Analysis and Prediction on Online Retail Sales

Sina Ainesazi Dovom
sinaainesazi@gmail.com

1.Introduction

The goal of this project is to apply Data Analysis, Data visualization, Modelling and Time-Series Analysis on an open source ‘Online Retail’ dataset.

I will introduce the dataset and explain the size, variables, and some information about the nature of the dataset and then I will explain the steps been taken for data preprocessing, data cleaning and exploration along with the plots achieved from the data visualization part and mention the interesting information gained from this step.

Also, we will discuss about the predictive modelling in which some approaches have been taken to build machine learning models for the Sales of the products and also the Time-series analysis which for this task the ARIMA models have been used which are one of the main useful models in forecasting tasks and will show the achieved results.

2.DataSet

The dataset that is used in this project, includes 541909 observations and 8 variables which are, 'InvoiceNo', 'StockCode', 'Description', 'Quantity', 'InvoiceDate', 'UnitPrice', 'CustomerID', 'Country'.

Generally, the rows are the contents of the invoices and each row has an InoiceNo, a variable that indicates the number of invoice which the product belongs in, a StockCode column that indicates the code for the product which has been assigned in the store, and a text column which is Description variable which containing the details and the name of the product, the Quantity variable that gives the information about the number of that product having purchased, InvoiceDate column which gives the information about the year, month, day and also hour in which the Invoice have been generated, UnitPrice column which tells the price for the product, CustomerId column which Customer number a Nominal, a 5-digit integral number uniquely assigned to each customer and we have Country name. Nominal, the name of the country where each customer resides.

3.Data Cleaning and Exploration

The first step was to check on the null values.

```
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID    135080
Country            0
```

Only two variables had missing values. Description and CustomerID, when I check on the rows containing missing values for the CustomerID, nothing special was observed and since this column had unusually high number of Nan values, I decided to drop it from the data. But, for the Description variable when I checked on the rows containing missing values for Description, it was strange because in the rows which the description variable is missing the UnitPrice is 0.0.

So, I can just fill in those rows for the Description variable as 'UNKNOWN'.

I also converted the InvoiceDate variable into DataTime.

I used ‘describe’ method to check on a quick summery statistics for "Quantity" and "UnitPrice" columns.

	Quantity	UnitPrice
count	541909.000000	541909.000000
mean	9.552250	4.611114
std	218.081158	96.759853
min	-80995.000000	-11062.060000
25%	1.000000	1.250000
50%	3.000000	2.080000
75%	10.000000	4.130000
max	80995.000000	38970.000000

The statistics showed a few interesting facts.

1. The minimum values for both 'Quantity' and 'UnitPrice' are negative, which is unusual.

2. We see that the standard deviations for both variables are quite large, suggesting a significant spread in the data. This, along with the massive gap between the max and 75th percentile, indicates the presence of outliers.

It might be a good idea to find out the reason, why do they contain negative values?

When I checked on the negative values in Quantity variable I observed that InvoiceNo starts with C, like this C536379. This

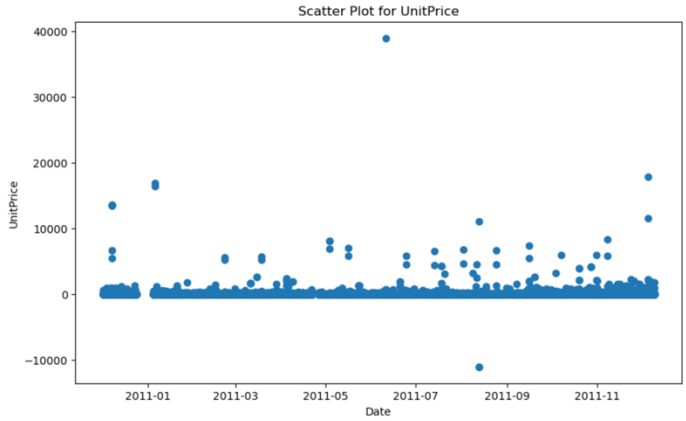
might be the word 'Cancel'. So, that's the reason why this column contains negative numbers.

What about UnitPrice? Only two rows were containing the negative number for the UnitPrice and in those columns the InvoiceNo was like this A563186, and the description was "Adjust bad debt", this could mean that the company is recording a transaction to adjust for a bad debt. We see this term is associated with negative values in UnitPrice, it could indicate that a sale was reversed because the debt was deemed uncollectable.

This often happens when a company decides to write off a bad debt. So, we see the InvoiceNo starts with 'A', this might be the word 'Accounting'.

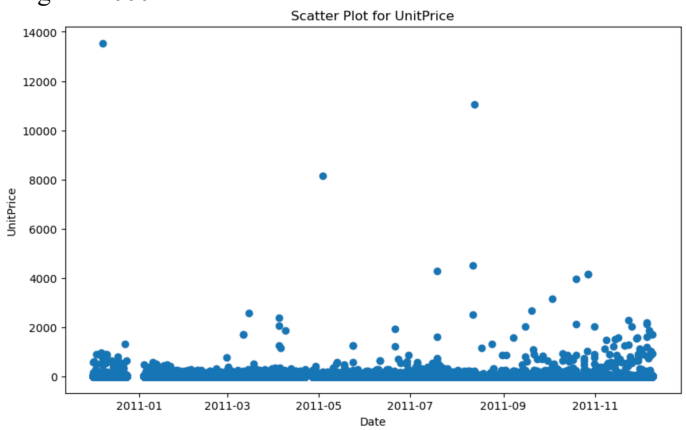
Next Step was to detect the outliers, It might be a good idea to inspect these distributions with scatterplot.

For UnitPrice:



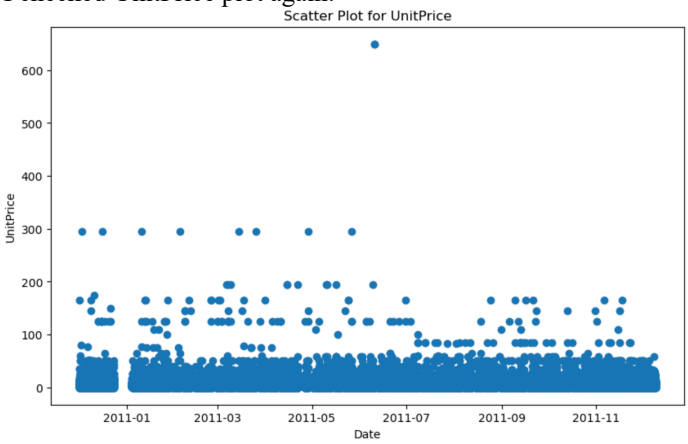
This plot shows a negative outlier for the UnitPrice which we found the reason of that and also one datapoint is extremely out of range which is close to 40K this also is an outlier so we should eliminate these from the data. So, I eliminated the negative values and very strange high values larger than 20K.

After this elimination, I noticed some strange outliers. In the range of 2000 to 14k.



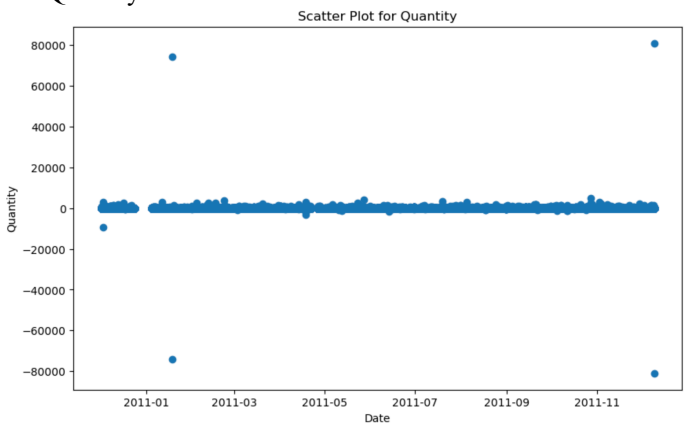
When I checked the columns with high UnitPrice, the detected outliers were 'AMAZON FEE', 'POSTAGE' and 'Adjust bad debt', 'Manual' and 'DOTCOM POSTAGE' and based on the purpose of our analysis, we won't get useful and helpful information from these. Therefore, eliminating these instances from the data would be reasonable.

I checked UnitPrice plot again:



This product was PICNIC BASKET WICKER 60 PIECES which has high price and in only 2 invoices contain this product and they're strange because in one invoice the quantity of this product is 1 but in the other it's 60. Despite being rare and high-priced, they are valid transactions. They could provide interesting insights, for example, demonstrating that high-ticket items, though infrequently purchased, can contribute significantly to sales. But since we will develop a predictive model to forecast sales, these outliers might distort our model. In this case, I might consider removing these data points to avoid biasing the models.

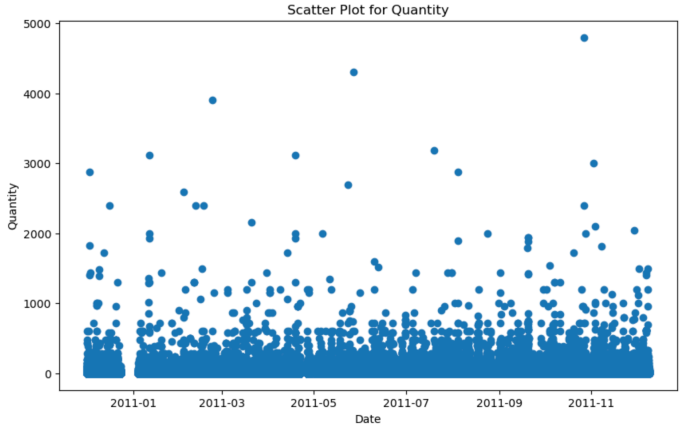
For Quantity:



This plot shows negative outliers for the Quantity variable which we can guess the reason and two datapoints are extremely out of range which are close to 80K, these also are outliers so we should eliminate from the data.

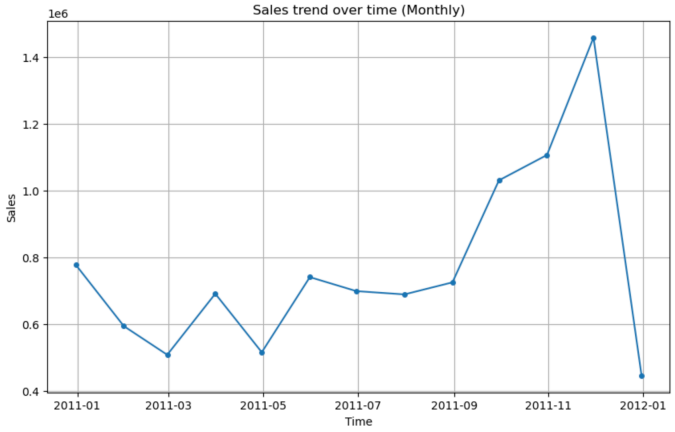
Actually, the negative values also could be a potential subject to go deep and analyze these kinds of Invoices but for this project I decided to ignore those.

I checked the scatterplot once more:



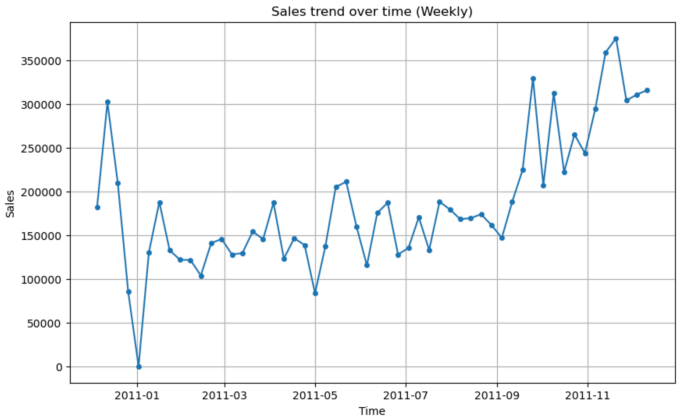
The products with the Quantity range of 2000 to 5000, I could not notice any strange thin in those transactions and I suppose they are valid.

For the visualization step, I created a new column representing the Sales, which is the product of the UnitPrice with the Quantity in each row, which shows the total cost for the product in that Invoice.
I checked on the Sales trend over time:



This plot shows the monthly Sales in all the countries, we see over all the trend is increasing from the end of the year 2010 to the end of 2011. But there's a huge drop in the beginning of the year 2012.

I also checked on the weekly Sales:

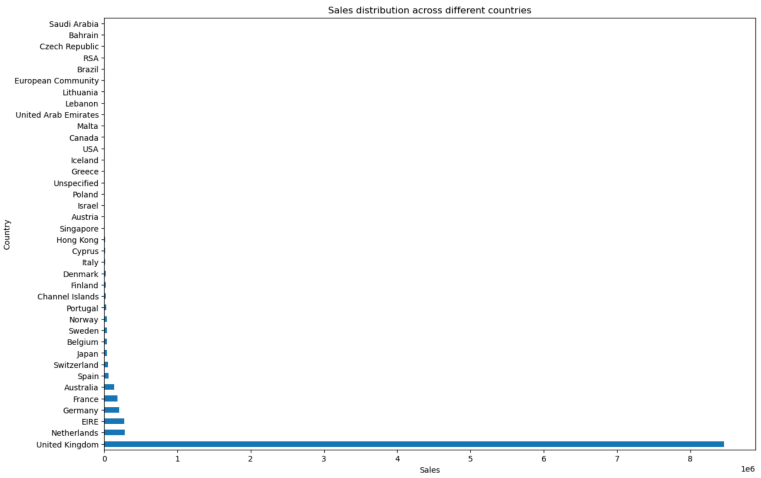


This trend also showed an overall increasing Sales but the strange issue from the plot observed. Why the sales are zero in 2011-01?! When I check on this occurrence,

InvoiceDate	
2010-12-05	181559.68
2010-12-12	302110.98
2010-12-19	209421.77
2010-12-26	85530.93
2011-01-02	0.00
2011-01-09	130595.67
2011-01-16	187678.06
2011-01-23	132518.83
2011-01-30	121738.30
2011-02-06	121544.43
2011-02-13	103757.97
2011-02-20	140867.03
2011-02-27	145593.66

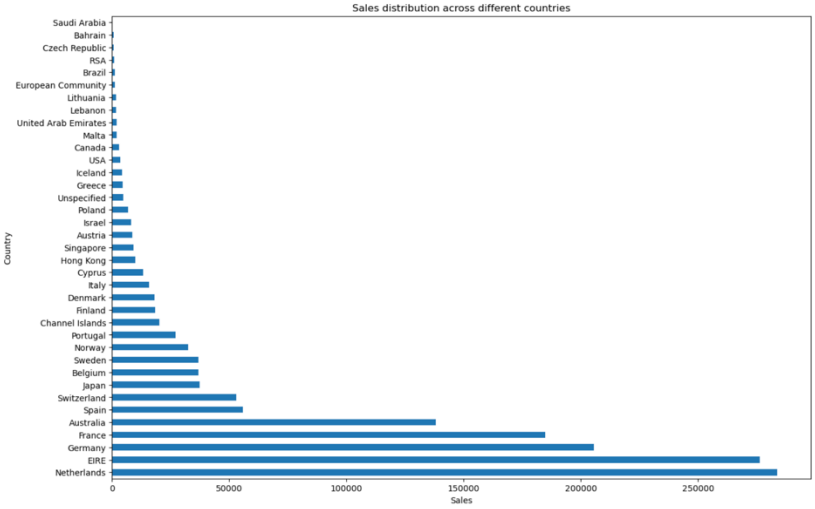
My conclusion for this issue is, maybe the store was closed in that week because of the new year. and there's no transactions have been occurred.

I checked on the Sales distribution across different countries:



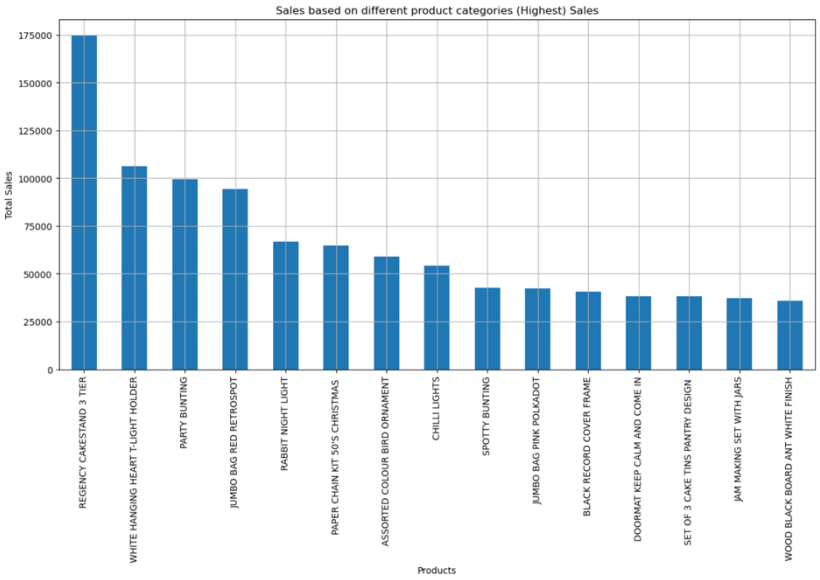
This barplot showed that the majority of the Sales is related to the UK, when I checked precisely, I observed that almost 92 percent of the Sales were related to UK in the dataset.

Let's see what countries after UK have the highest Sales:

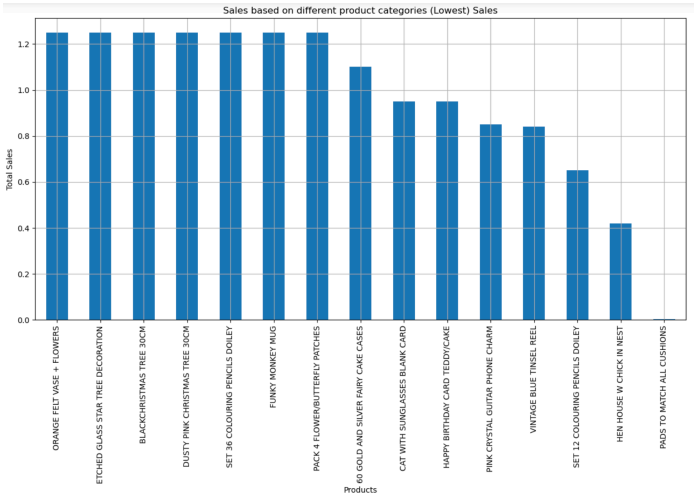


After UK, top best countries are Netherlands, Ireland, Germany, France, and Australia.

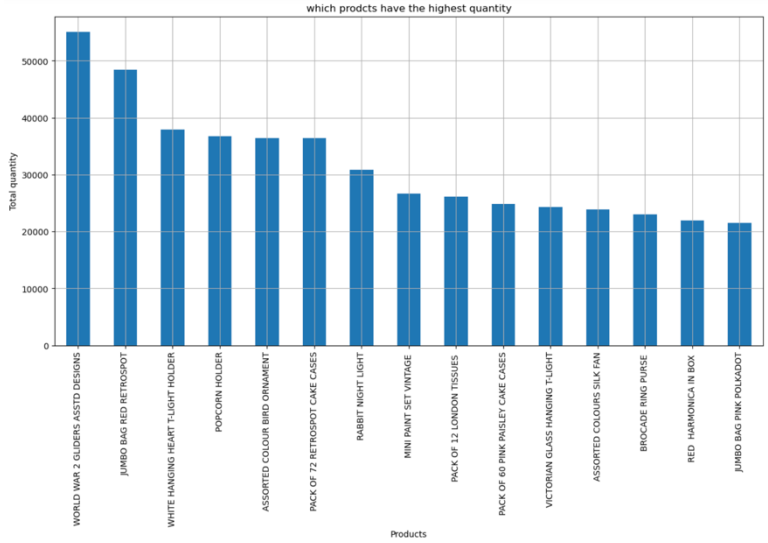
I checked on the top 15 products with highest Sales.



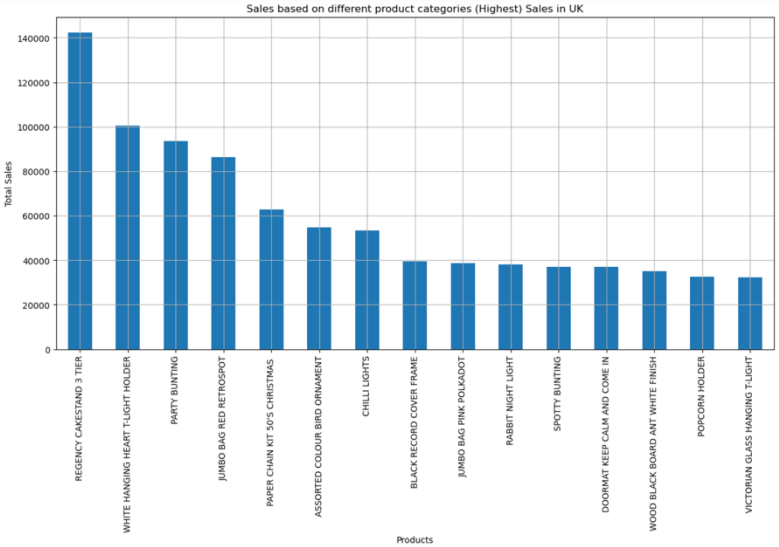
And the 15 products with the lowest Sales:



Let's see which products have the highest Quantity in the Invoices:

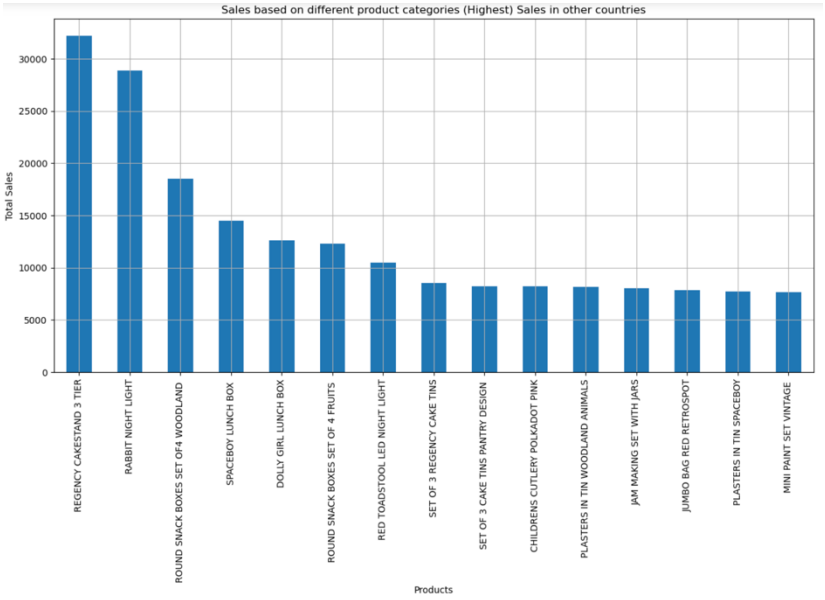


Which products in the UK have the highest Sales, this is very important factor since we have seen the 92 percent of the Sales



are for UK and finding out the products having highest Sales is crucial:

And also, which product have the highest Sales in countries other than the UK, This might help the store to see which product are popular and have high Sales and try to keep them available in the countries:



The 'Regency cakestand 3 tier' is the product which this store has its highest Sales, both in the UK and other Countries.
Of course we could do so many Data visualization and find so many valuable information and factors but these were the most interesting facts that we could gain from this data.

4.Predictive Modelling

We are going to build Machine learning models for the Analysis of Sales, and we are going to be using only the UK dataset for this purpose. Since UK has the highest number of sales among the other countries.
For Modelling part 2 approaches have been done

1. Modelling for the product sales.
2. Modelling for the UnitPrice with new variables over time.

1.Feature Engineering Part

I extracted some more useful information like Year, Month, Day, and hour from InvoiceDate feature.

StockCode feature was type object and it contained int and str together so I had to handle this feature, I tried to apply one encoding strategy on this variable:

Which is Label Encoding: Using label encoding, where each StockCode is assigned a unique numeric identifier. This can be a good option for tree-based models, which can handle categorical features encoded in this way. I used the Lable Encoding before the splitting the dataset.

2.Modelling

I used LabledStockCode, Quantity, UnitPrice, Year, Month, Day, Hour to train the models and predict the Sales.

I Splitted the dataset into Training and Testing sets and I used the StandardScaler to fit_transform the training and testing sets. After these I adjusted the parameters for Linear Regression, Decision Tree and Random Forest and using the gridsearch cross validation I tried to fit the models on the data, the results for this part of the analysis is interesting because linear regression had the weakest performance in comparison to the DT and RF. Results are as follows:

```
 === Start report for regressor LinearRegression ===
Tuned Parameters: {'fit_intercept': True}
Best score is 0.4340521801753 44
MAE for LinearRegression
9.584676719467357
MSE for LinearRegression
2246.3321466554703
R2 score for LinearRegression
0.451872233580899
 === End of report for regressor LinearRegression ===

 === Start report for regressor DecisionTreeRegressor ===
Tuned Parameters: {'min_samples_leaf': 2, 'min_samples_split': 2}
Best score is 0.9622217795775703
MAE for DecisionTreeRegressor
0.2872679274776991
MSE for DecisionTreeRegressor
177.93813940708262
R2 score for DecisionTreeRegressor
0.9565812940623274
 === End of report for regressor DecisionTreeRegressor ===

 === Start report for regressor RandomForest ===
Tuned Parameters: {'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 100}
Best score is 0.9746236821601209
MAE for RandomForest
0.1760232459959917
MSE for RandomForest
70.26949800540703
R2 score for RandomForest
0.98285353167988
 === End of report for regressor RandomForest ===
```

The second approach for modelling was, Unitprice analysis over time. And for this approach I tried to to add other variables and to bucket them.

I made a new variable QuantityInv which indicates the total quantity of products in each invoice. Then I checked on the distribution of UnitPrice and Quantity in the Data to help me to make intervals and bucket the data.

Again, I extracted month feature from the Date variable and added to the dataset. Again, I splitted the dataset by countries to use UK dataset only.

I created dummy variables from the variables I had created and bucketed. And I scaled the QuantityInv variable to has scale close to the other variables in the dataset.

I splitted the dataset like before into Training and Testing sets and also using the previous machine learning models and adjusting their related parameters I tried to fit them on the data and the result on the test set are as follows:

```
 === Start report for regressor LinearRegression ===
Tuned Parameters: {'fit_intercept': True}
Best score is 0.6245911028736323
MAE for LinearRegression
0.9223954212476825
MSE for LinearRegression
9.108545933390449
R2 score for LinearRegression
0.5687136914693902
 === End of report for regressor LinearRegression ===

 === Start report for regressor DecisionTreeRegressor ===
Tuned Parameters: {'min_samples_leaf': 2, 'min_samples_split': 2}
Best score is 0.5957542730531629
MAE for DecisionTreeRegressor
0.8299825016778437
MSE for DecisionTreeRegressor
7.675146477828906
R2 score for DecisionTreeRegressor
0.6365846298562388
 === End of report for regressor DecisionTreeRegressor ===

 === Start report for regressor RandomForest ===
Tuned Parameters: {'min_samples_leaf': 2, 'min_samples_split': 3, 'n_estimators': 100}
Best score is 0.6431029905550633
MAE for RandomForest
0.8250516872181533
MSE for RandomForest
7.417054361681814
R2 score for RandomForest
0.6488051968762594
 === End of report for regressor RandomForest ===
```

We can see that the Linear regression for this kind of analysis had better result and the best model with good results of MAE and MSE and R2 is for Random Forest.

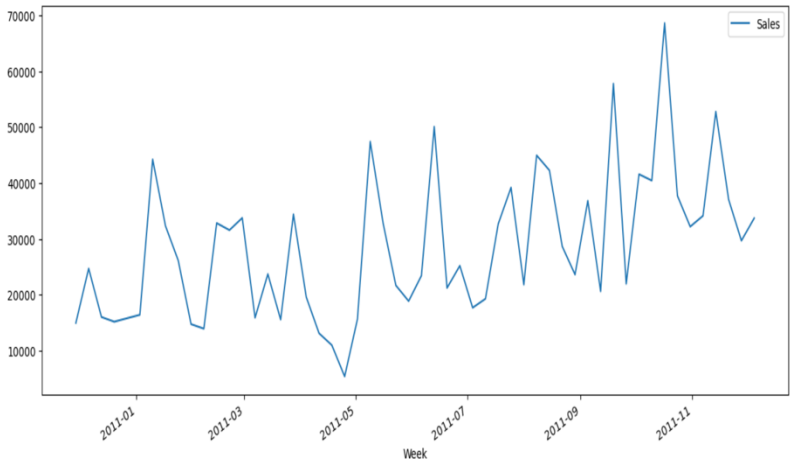
3.Time-Series Forecasting

For this task I decided to use dataset for countries other than UK. And I used ARIMA models to forecast the next 6 months.

We are interested in forecasting the next 6 months. but since monthly data gives us a very short series I will be using the weekly data to have longer serie and try to check on the series to see if it's stationary or not and whether the serie needs some differencing.

I made the weekly time series for sales.

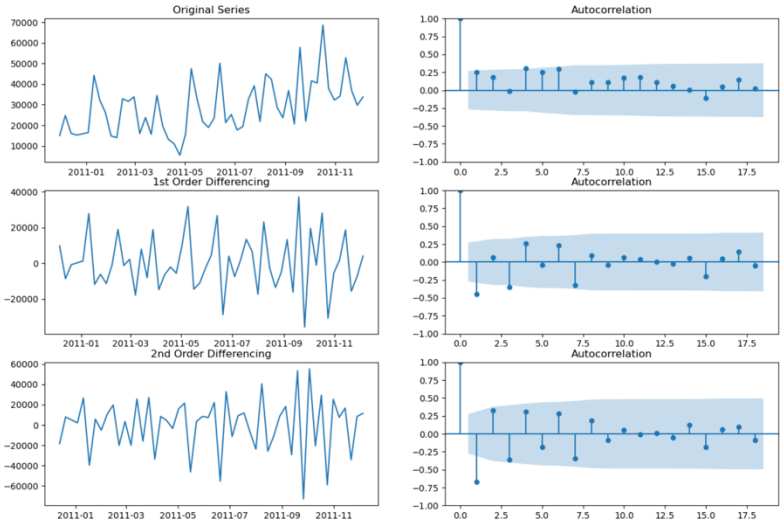
I check on the weekly Sales:



This Plot shows that the series is not stationary and follows a trend. To check more on this issue, I used adfuller test and the result was:

```
ADF Statistic: -2.234714
p-value: 0.193888
Critical Values:
```

The P-value is greater than 0.05 and this shows that the series is not stationary and needs differencing. I applied two levels of differencing and the results are as follows:

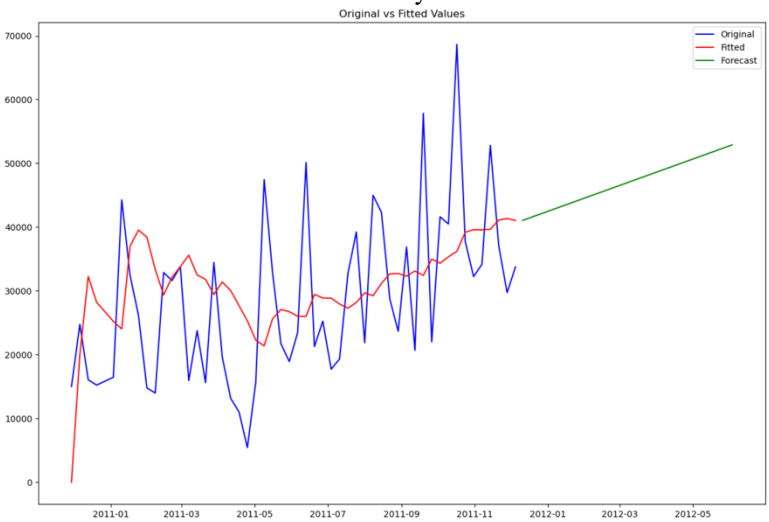


This shows that adding 1 differencing makes the series stationary.

I tried to generate all different combinations of p, d and q triplets for ARIMA models and compare the BIC and AIC parameters to receive the model with the lowest parameters among the others. The result was this:

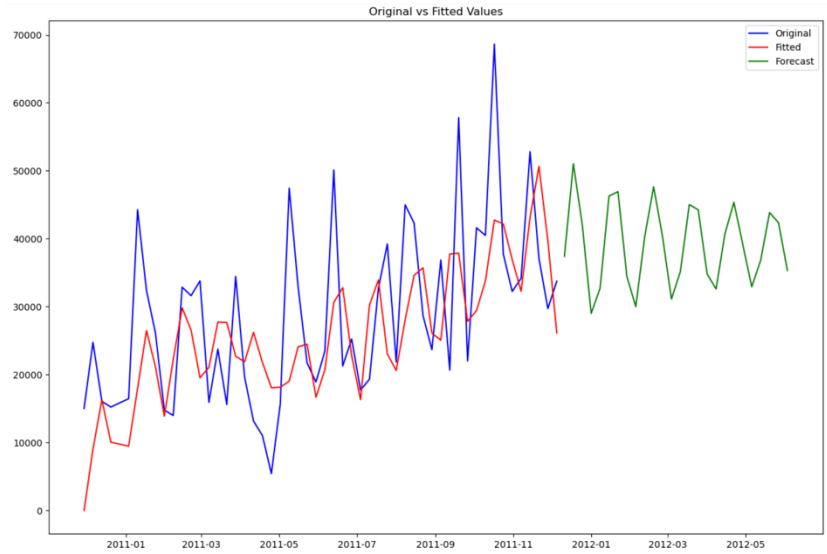
```
The best ARIMA model is ARIMA(0, 2, 2) wi
th BIC=1124.3363460005053
The best ARIMA model is ARIMA(0, 2, 2) wi
th AIC=1118.5408691023324
```

However, when I tried to visualize this model when it fits and forecasts the result was not satisfactory:



It's obvious that it cannot capture the trend and when it tries to forecast the next months it's only a linear increasing line.

I tried to find another model manually and I reached the ARIMA model with order (3,1,3) and I tried to visually see it:



This model can capture the trend, up and downs of the series better and also the forecasting shows more satisfying results that kind of is similar to the historical data. However I splitted the data into training and testing and checked on the performances of both of the models and for the ARIMA with order (3,1,3) the results are

```
MAE: 10330.256
RMSE: 13163.232
```

And for the ARIMA model with order (0,2,2) the results are:

```
MAE: 10417.951
RMSE: 13441.392
```

Which shows the ARIMA(3,1,3) has better results in comparison to the other model.

5.Conclusion and recommendations

1. Product Sales in Various Regions: The UK accounts for approximately 92% of all sales. Other significant markets include the Netherlands, Ireland, Germany, France, and Australia. To further boost sales, the company should consider strengthening its marketing strategies in these areas.

2. Product Focus: The 'Regency cakestand 3 tier' is a top-performing product in both the UK and other countries. The company could consider running special promotions or deals on this product to encourage further sales. Additionally, understanding why this product is so popular could help in the development of future products.

3. Sales Fluctuations: There are noticeable fluctuations in the weekly sales data, with sales dipping to zero in early 2011, likely due to the store closure. The company may want to consider alternative strategies during these periods, such as online sales or pre-holiday promotions.

4. Addressing Negative Values and Outliers: There are negative values and outliers in both 'Quantity' and 'UnitPrice' data that may distort the analysis. These anomalies may be a result of transactions being canceled or due to bad debts. It's recommended to investigate these anomalies further to understand their root cause and take appropriate action to minimize their occurrence in the future.

5. Price Analysis Over Time: The analysis reveals a trend in unit prices over time. This insight can be used to predict future sales and develop pricing strategies. Additionally, the company might want to investigate the cause of price fluctuations to further optimize pricing.

6. Time-Series Forecasting: Utilizing ARIMA models for time-series forecasting could help the company understand sales patterns over time and make accurate predictions for future sales. This could assist in better inventory management and planning marketing strategies.

7. Inventory Management: A thorough analysis of the product sales can help in maintaining the right inventory. For example, ensuring that the top 15 products with the highest sales are always stocked can help increase the overall sales performance.

8. Product Variety: While focusing on top-performing products, it's also important not to ignore the lowest-selling products. Sometimes, it's the variety of products that attract more customers. Strategies can be developed to increase their sales, like bundling them with top-selling products.

References

[1] https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/?utm_content=cmp-true

[2] https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

[3] Pattern Recognition and Machine learning, Christopher M. bishop,2006. https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognitionand-Machine-Learning-2006.pdf