1. Objective

The objective of this document is to identify the important e-stores so that GloboSales can have an improved focus and retain their customer base.

For achieving this, e-stores have been analyzed and which have been generating more revenue by geography and their top customers. In addition, a list of the e-stores individually for GloboSales to concentrate on have been generated. The assumption being the more the revenue being generated, the more focus/concentration on the e-store/area. The analysis includes provides list of e-stores to focus on and predicts revenue for the top e-stores so that GloboSales can have the ability to make smarter business decision for future growth and expenses if any.

Given below are the key tasks that have been performed as part of the Analysis:

1. EDA – An exploratory analysis has been done to look at the data and identify data quality issues.
2. Top N Analysis – Analysis on finding the key e-stores and the customers contributing to the revenue
3. Forecasting – Simple forecasting for the top three e-stores individually and e-stores by Geography.

The entire analysis has been done in Python and includes 3 python files related to the above. The files have been uploaded into a GIT repository the link to which is at the end of the document. The csv data provided has also been uploaded into the same repository for the python files to read.

2. EDA

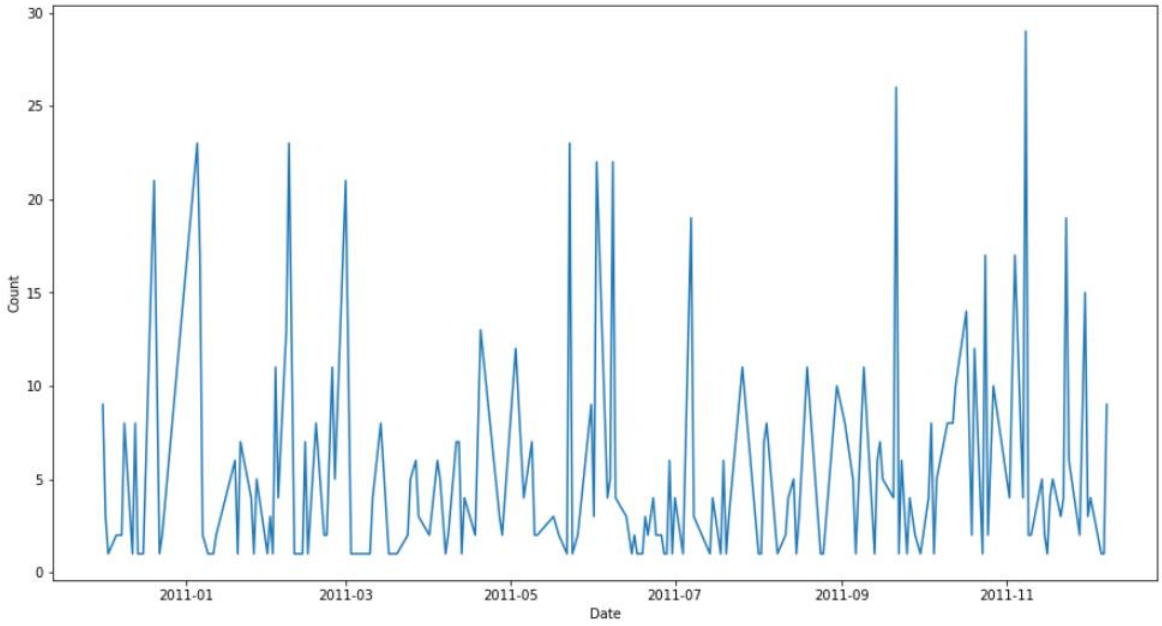
Following the exploratory analysis performed, below are the key points that have been identified:

1. There are 32 e-stores in total which serves around 1200 customers based in different geographies. A customer can be serviced by multiple e-stores.
2. There are 2 date field duplicates i.e., InvoiceDate of which one has been removed for analysis.
3. Removed ".0" in CustomerID column for all below analysis as this makes no sense since it is a string field.
4. 130000 rows and 10 columns for Dates between 2010-12-01 and 2011-12-09 were provided, after removing the duplicate Invoice Date column.
5. Around 120k nulls exist for columns ***Description, Quantity and StockCode*** as seen below.

```
###Count of nulls by Column
print(df_ecomm.isnull().sum(axis=0))   # Around 120k nulls in Stockcode,desc and quantity. No nulls in others
```

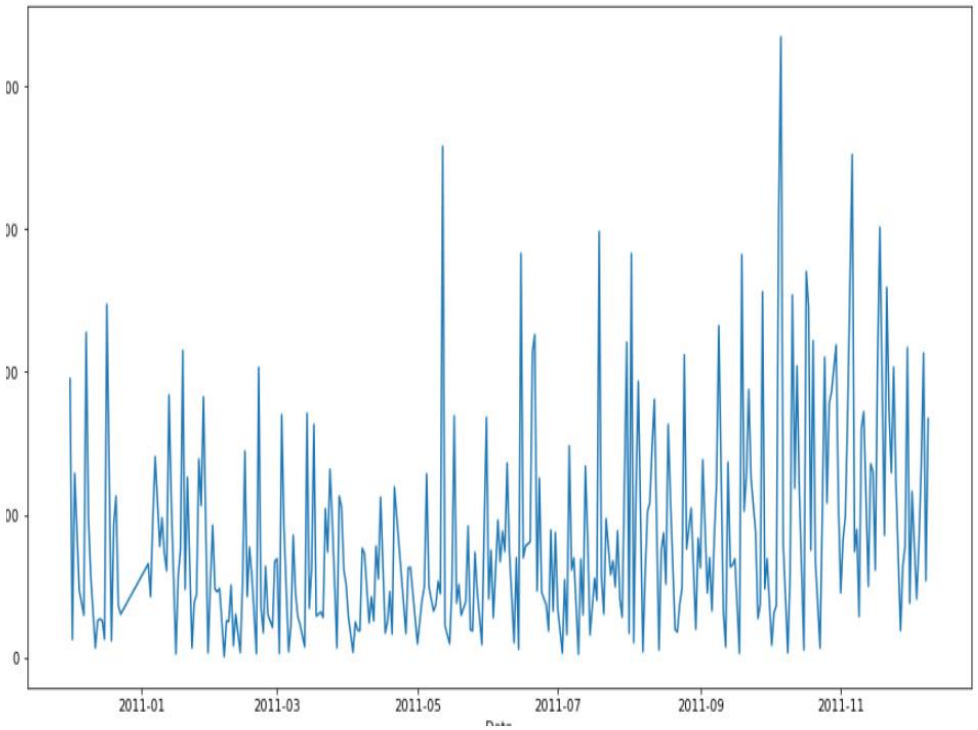
```
InvoiceDate         0
InvoiceNo           0
StockCode      123354
Description    126307
Quantity       119574
UnitPrice           0
CustomerID          0
Country             0
Estore_id           0
Date                0
dtype: int64
```

6. Negative Values in Quantity Column are 943 and in UnitPrice Column are 0 which are numeric features.



7. Most transactions are on 2011-Oct-06 and least on 2011-Feb-06 respectively

	Date	Count
249	2011-10-06	2174
125	2011-05-12	1791
275	2011-11-06	1762
248	2011-10-05	1520
286	2011-11-18	1507
...	...	...
233	2011-09-18	14
60	2011-02-20	13
30	2011-01-16	12
174	2011-07-10	11
48	2011-02-06	3

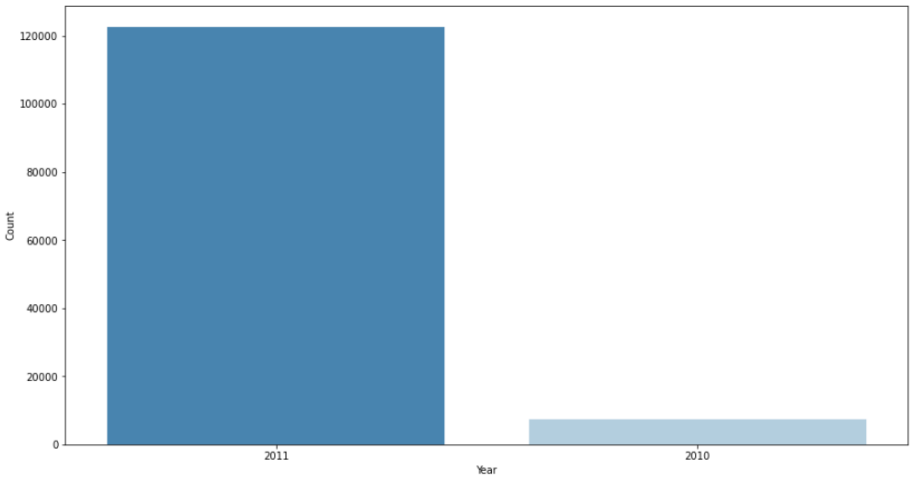


8. Unique values by columns after removing nulls, with 1920 unique Invoice numbers and 1246 Customers

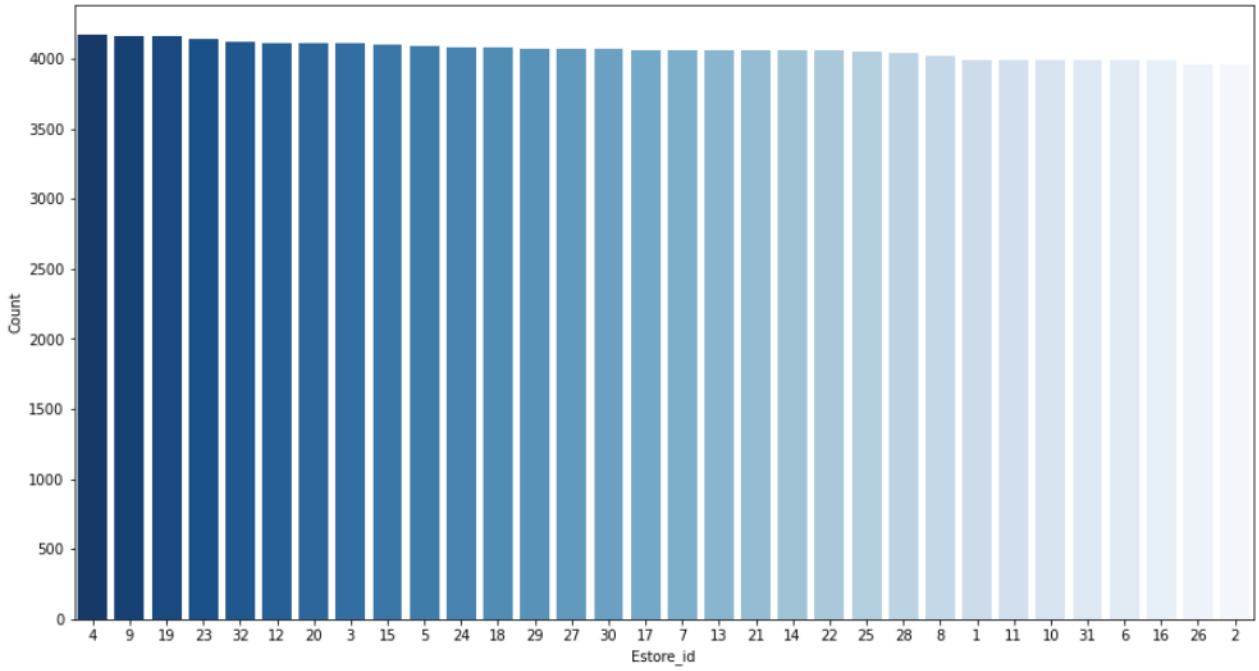
```
: ###Unique values by column excluding Nans

  df_ecomm.dropna().nunique()
```

```
: InvoiceDate    1907
  InvoiceNo      1920
  StockCode      1168
  Description    1184
  Quantity         70
  UnitPrice       123
  CustomerID     1246
  Country           7
  Estore_id        32
  Date            302
  dtype: int64
```

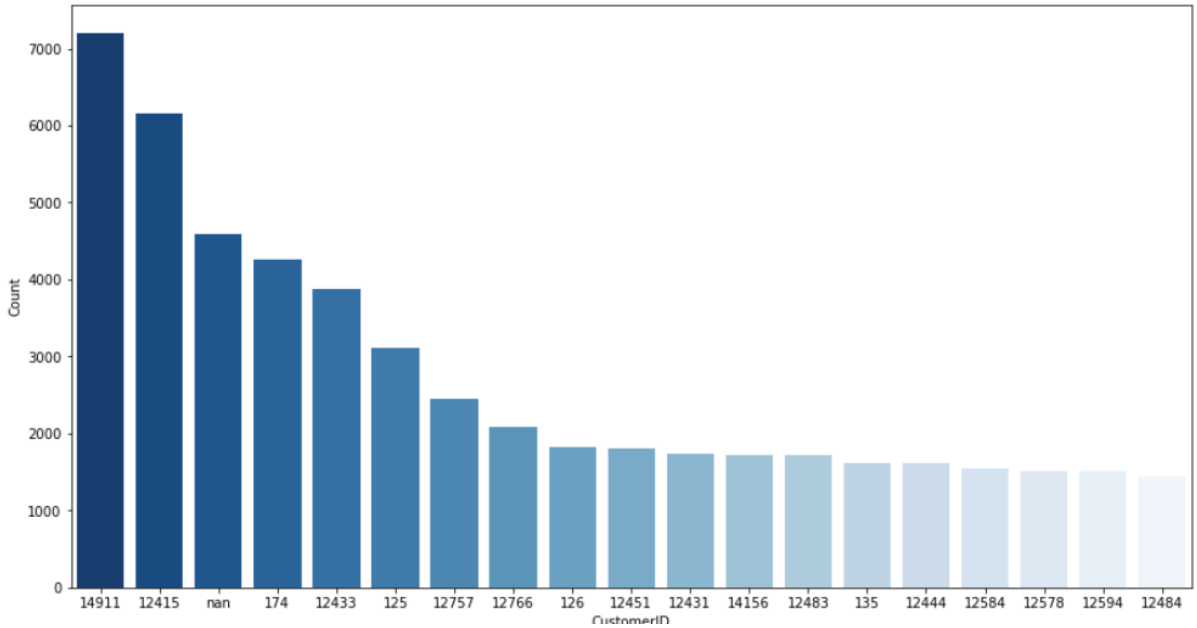
9. Transactions by Year . All transactions shown below with negative quantity values and nulls included.



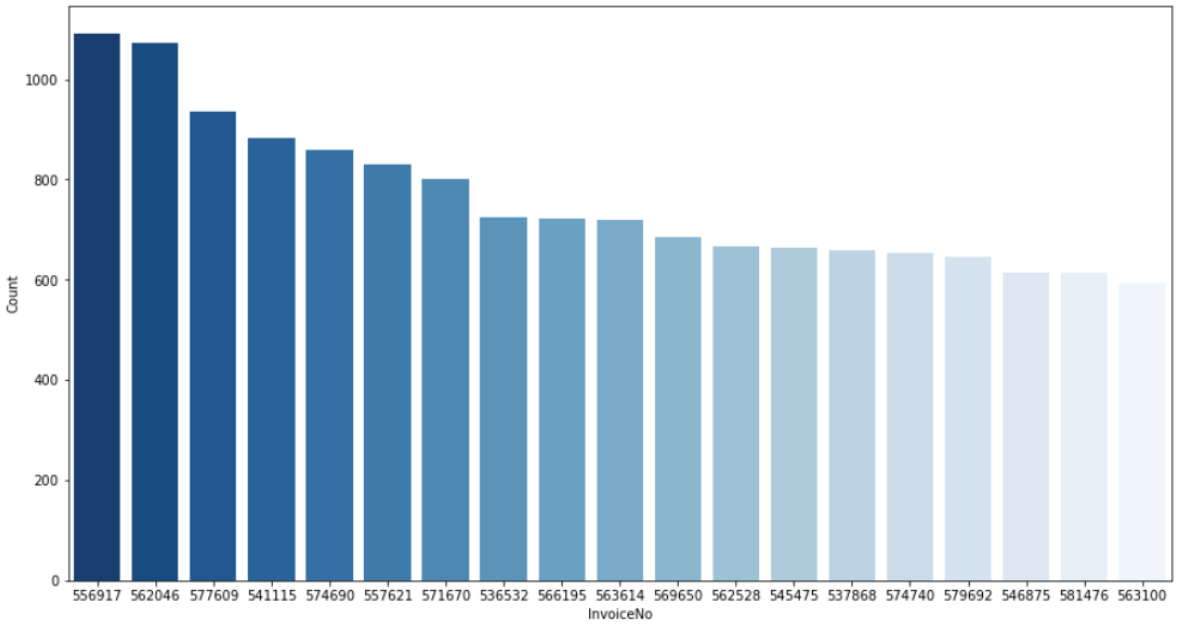
10. Count of Transactions by E-Store



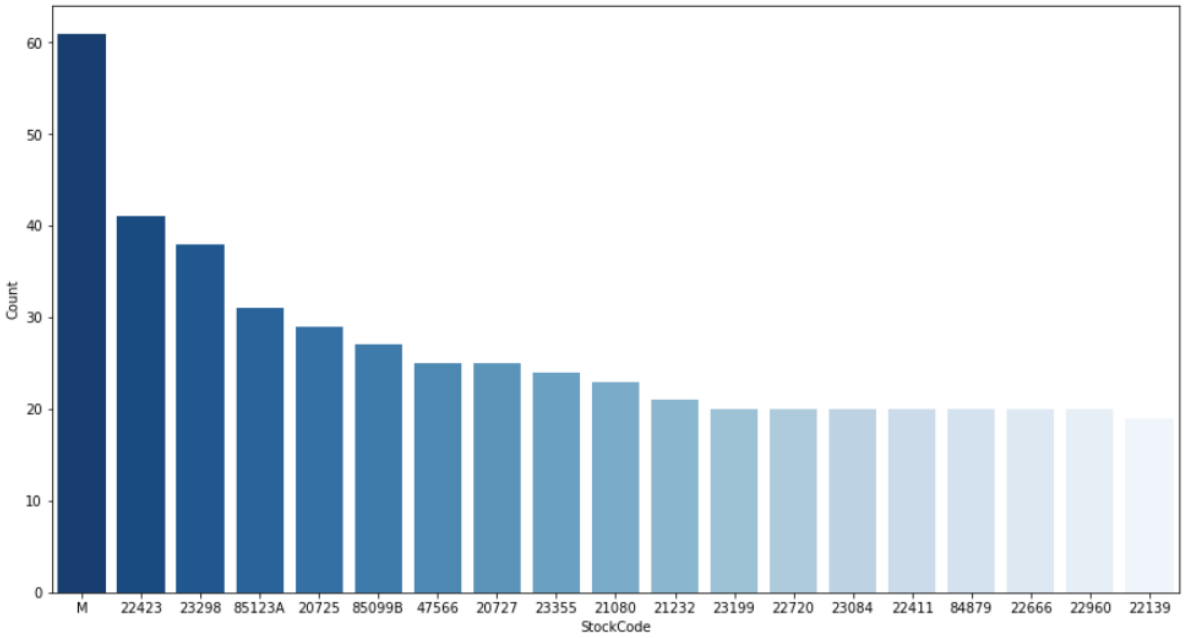
11. Top 20 Customers by Transaction Count. As seen below null counts are third in order.



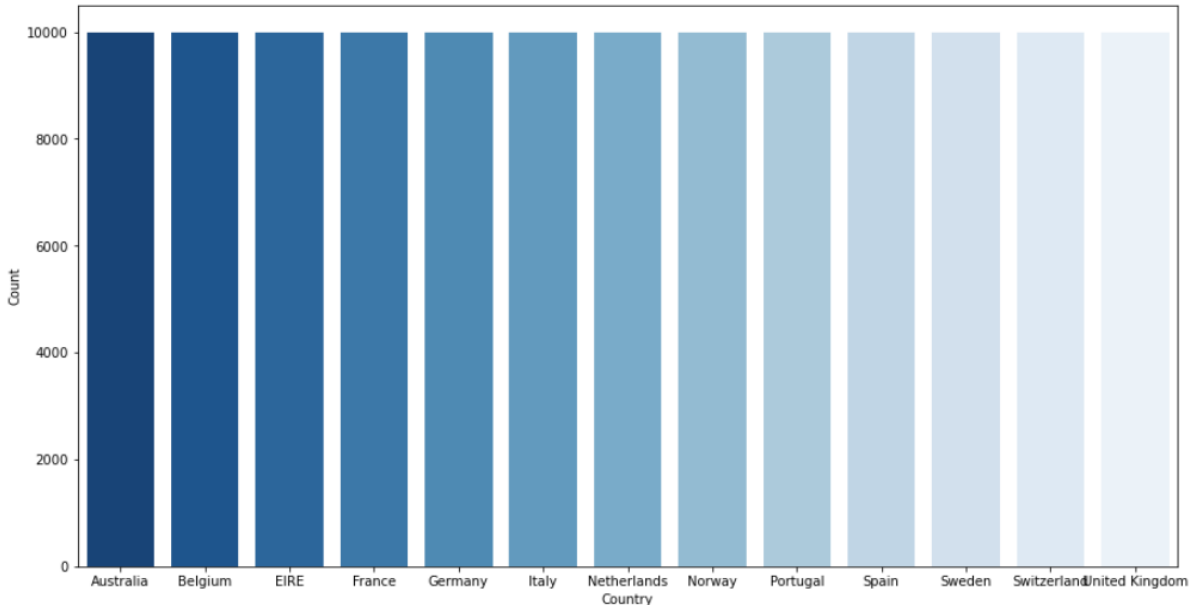
12. Top 20 Invoice numbers by Transaction Count



13. Top 20 StockCodes by Transaction Count



14. Count of Transactions by Country



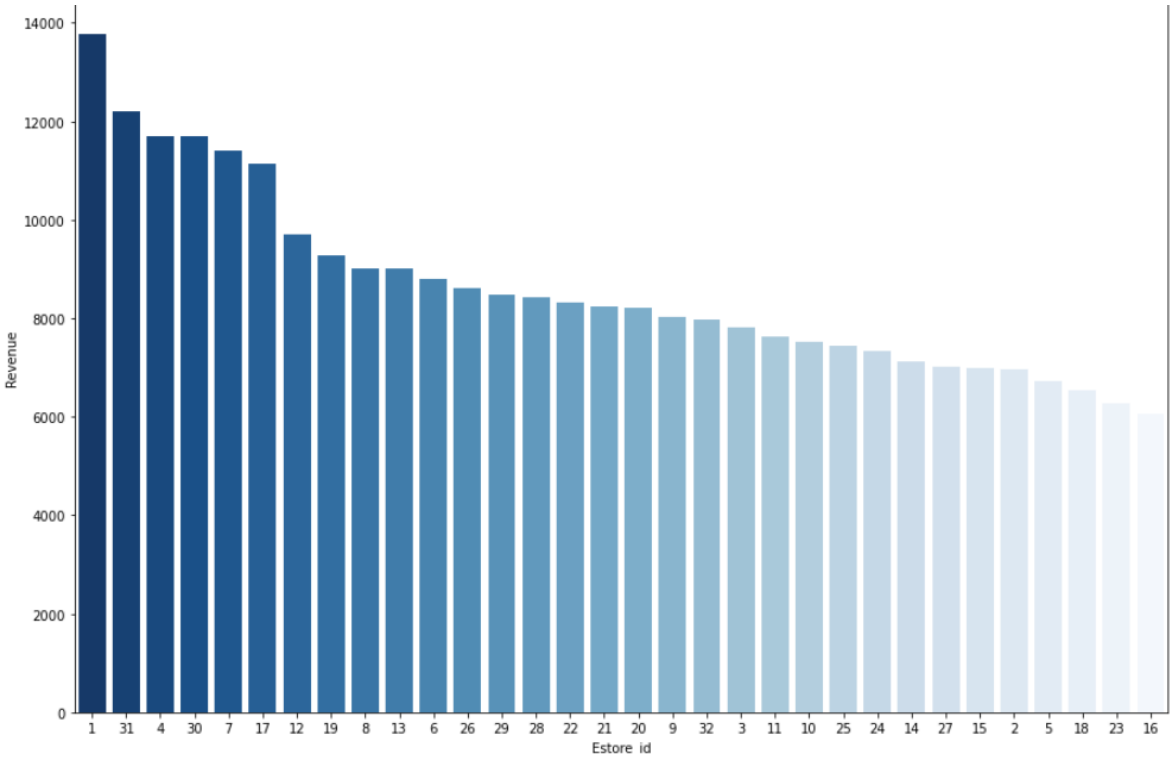
3. Top E-stores by Revenue

Top E-stores contributing to revenue have been identified instead of transactions as all the e-stores have similar number of transactions i.e., invoices as seen in point 10 above. This holds true for the forecasts as well. The revenue feature generated is the product of Quantity and Unit Price.

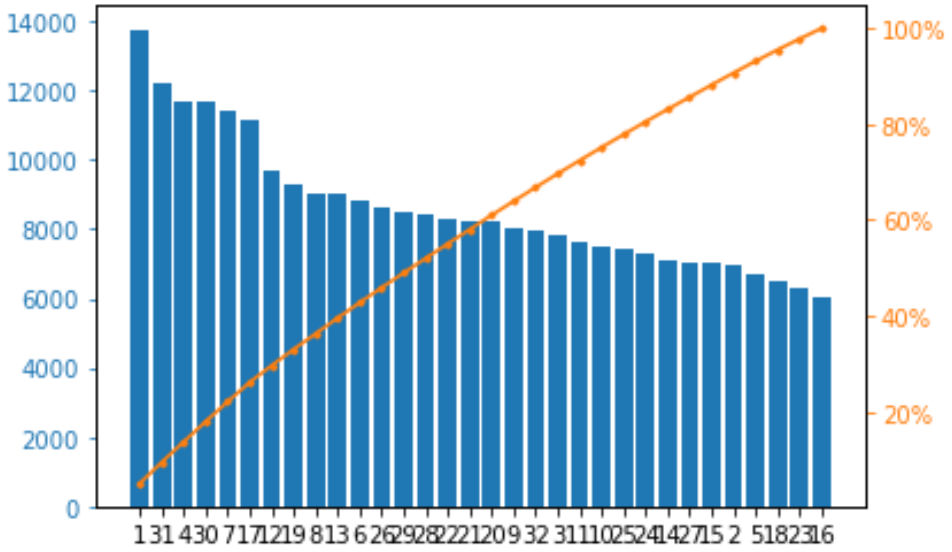
All the rows with negative quantity values have been removed along with the nulls. The top e-stores having the most amount of revenue are identified along with the top countries having the most revenue and their associated e-stores and customer IDs. In the forecast analysis below as well, we look at the top e-stores.

The recommendation is to focus on these stores as way of increasing revenue and also on the bottom five to see why the e-stores have lesser revenue even while having similar number of transactions.

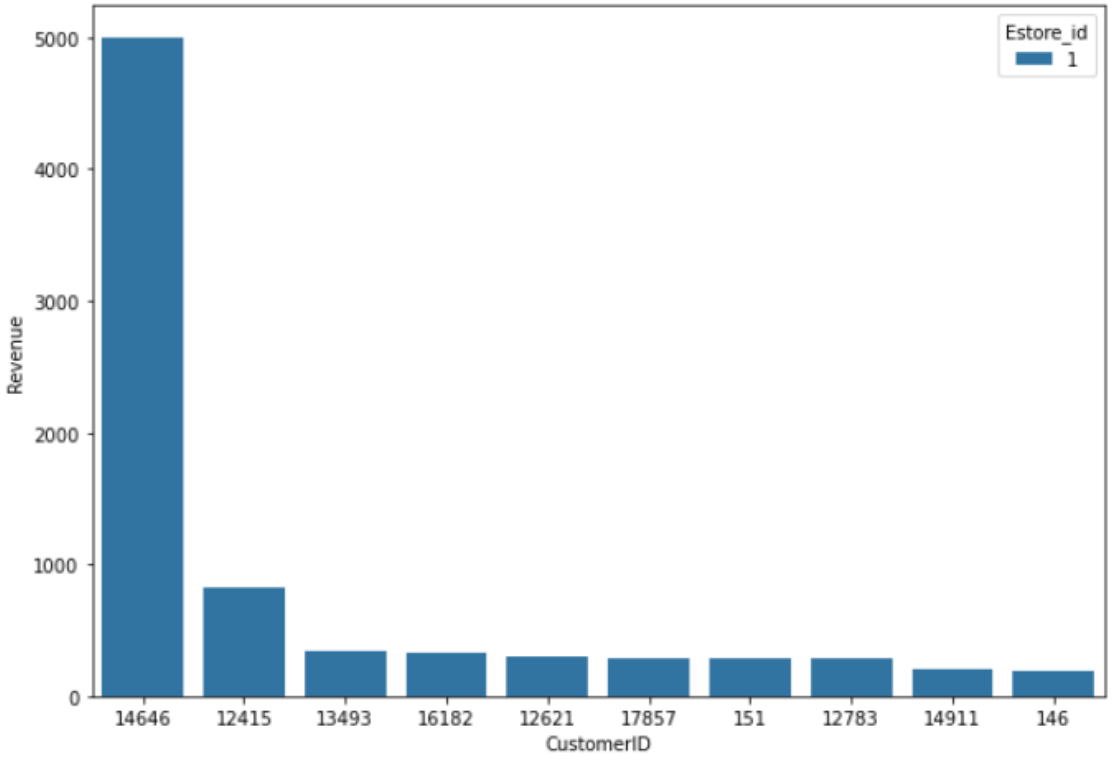
1. The top contributing five e-stores by Revenue are e-stores **1,31,4,30,7.**

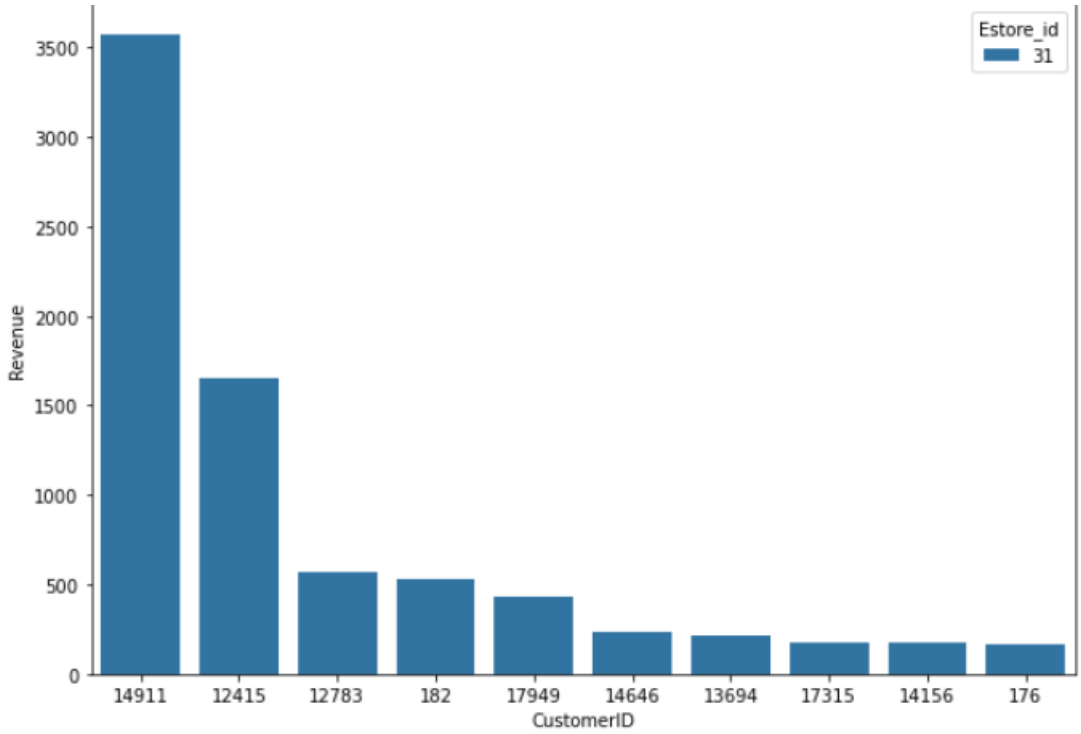


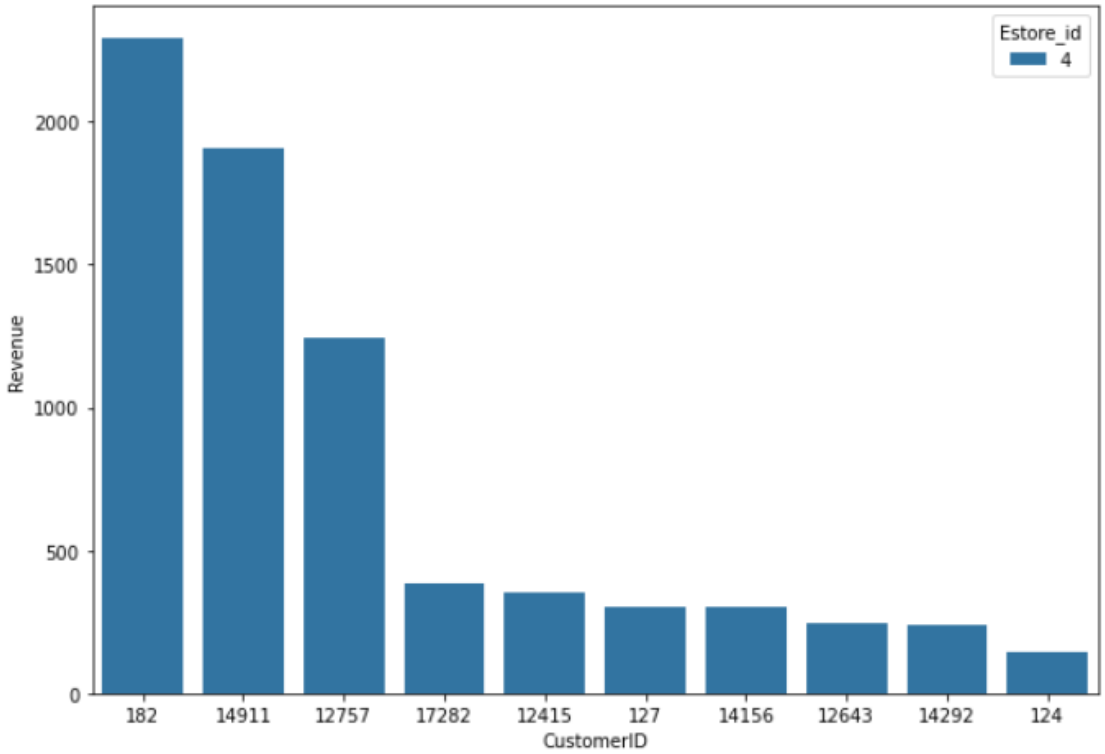
2. Almost all stores contribute to 80% of revenue barring stores **14, 27, 15, 2, 5, 18, 23, 16**. The below pareto chart shows the same.



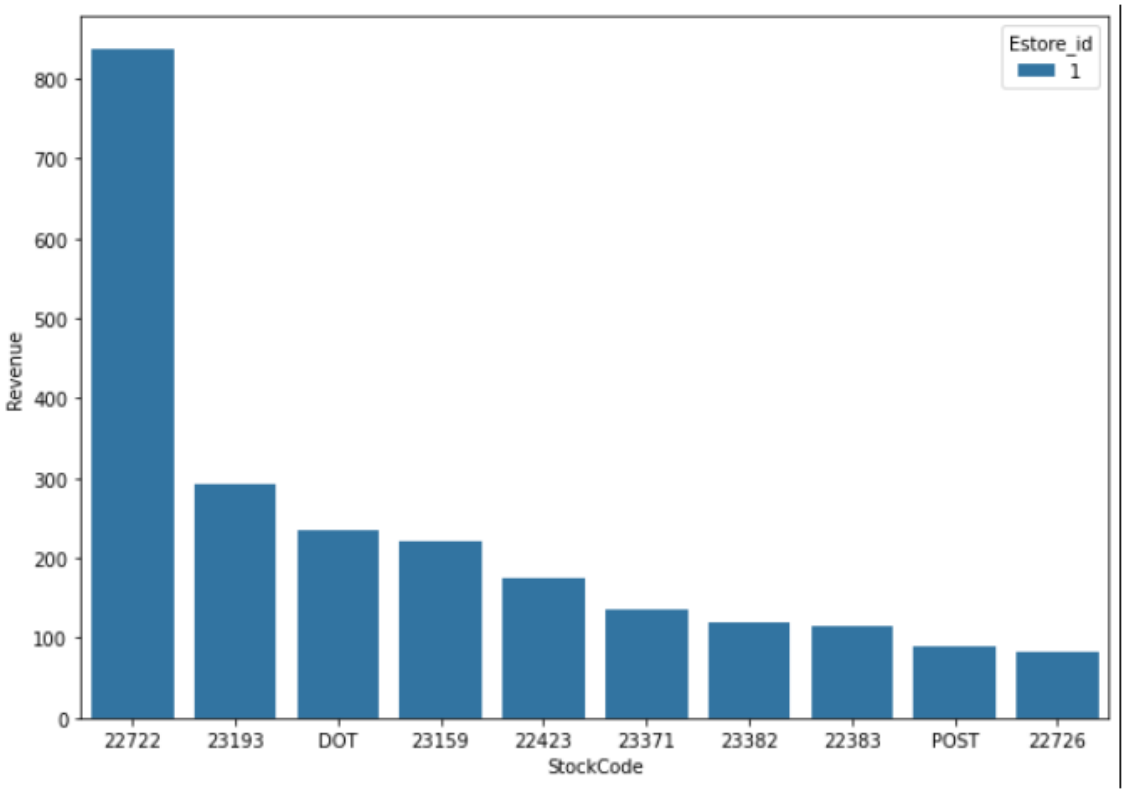
3. The top customers associated with the top e-stores **1,31,4** are as below:

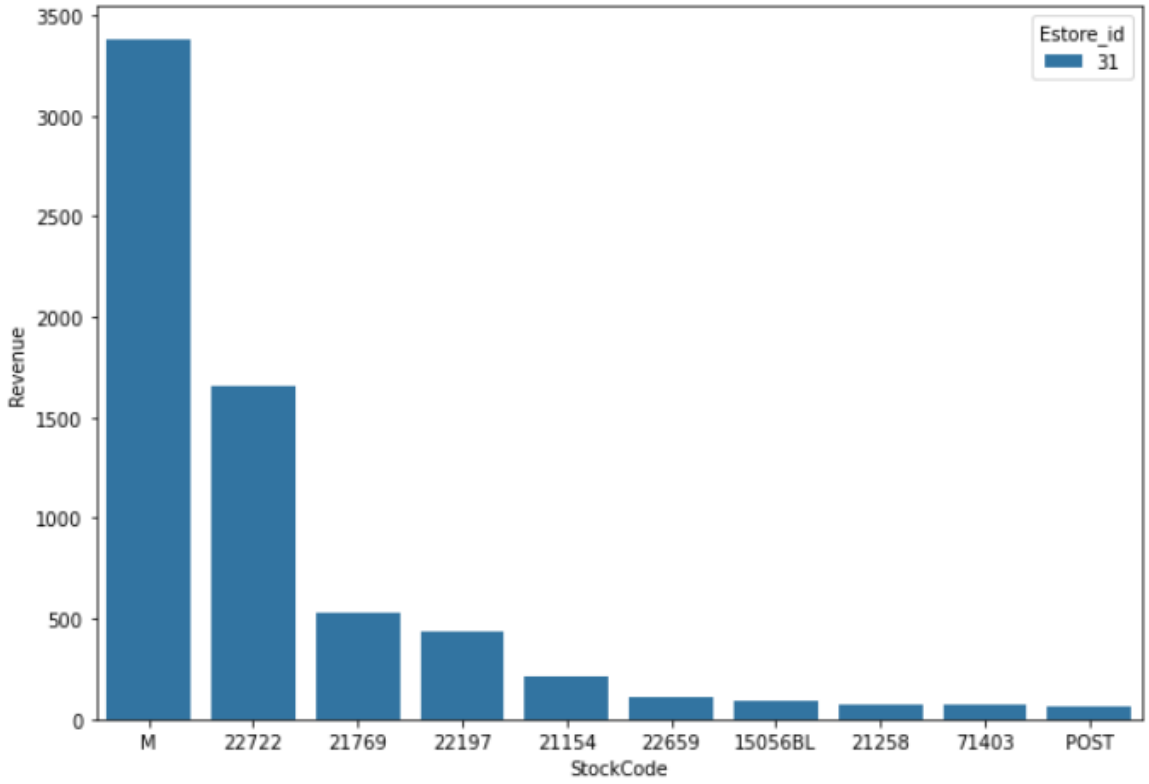


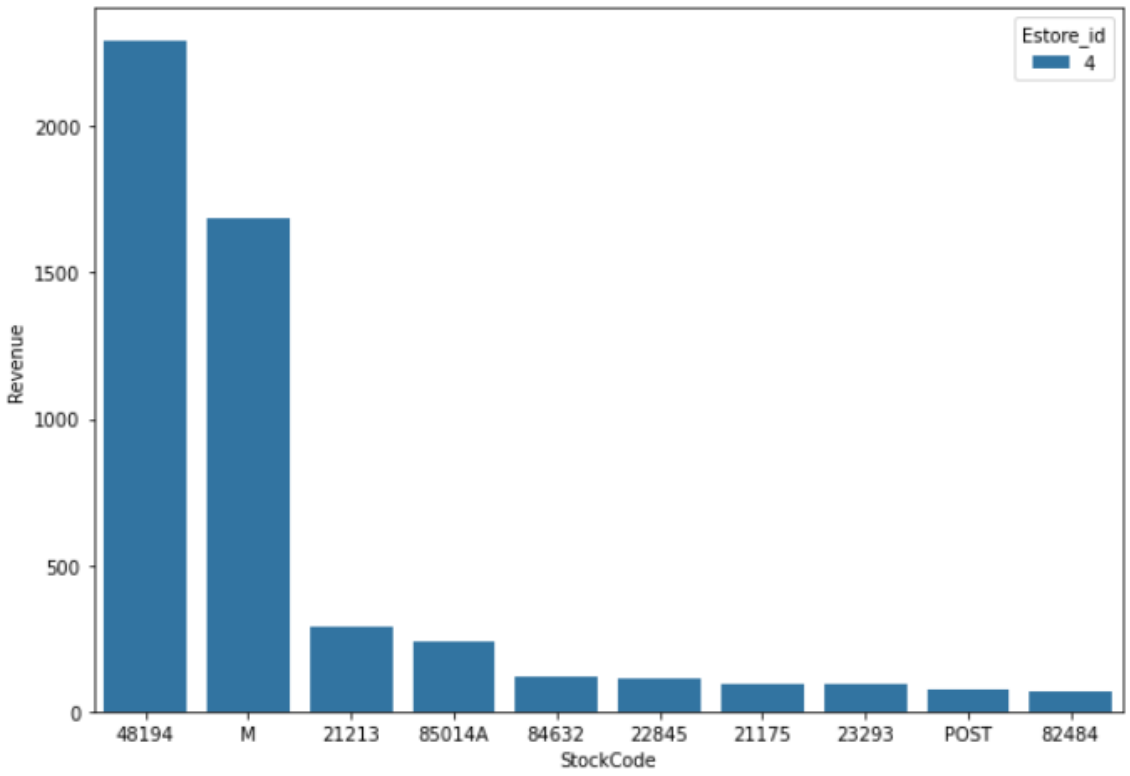

Estore_id
31
Revenue
3500
3000
2500
2000
1500
1000
500
0
14911
12415
12783
182
17949
14646
13694
17315
14156
176
CustomerID


Estore_id
4
Revenue
2000
1500
1000
500
0
182
14911
12757
17282
12415
127
14156
12643
14292
124
CustomerID

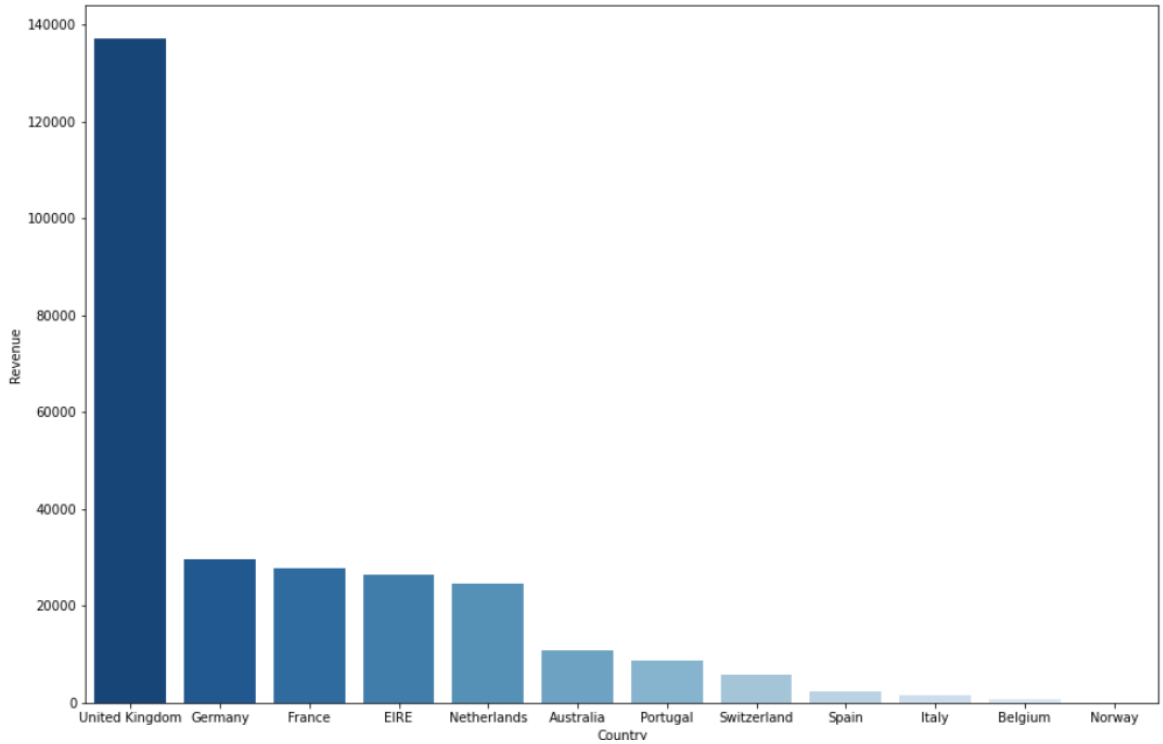
4. The top customers associated with the top e-stores **1,31,4** are as below:



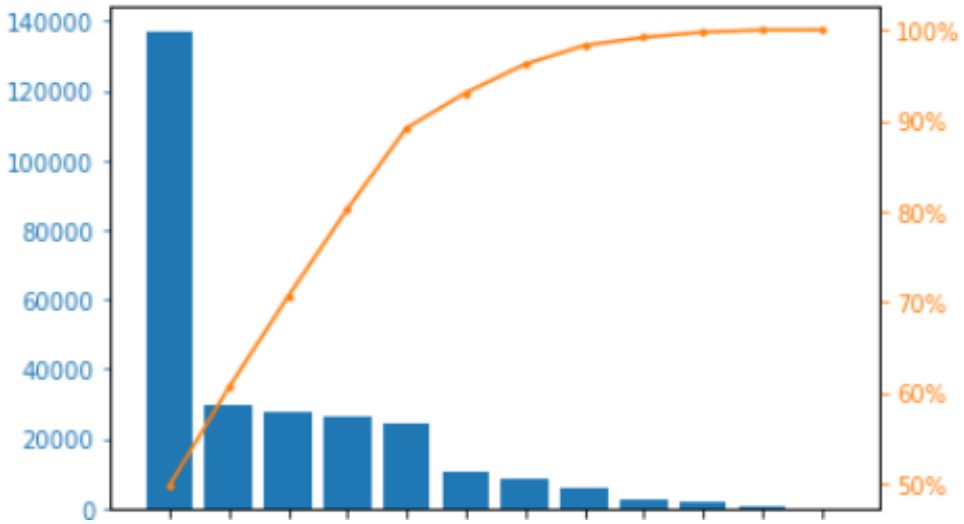




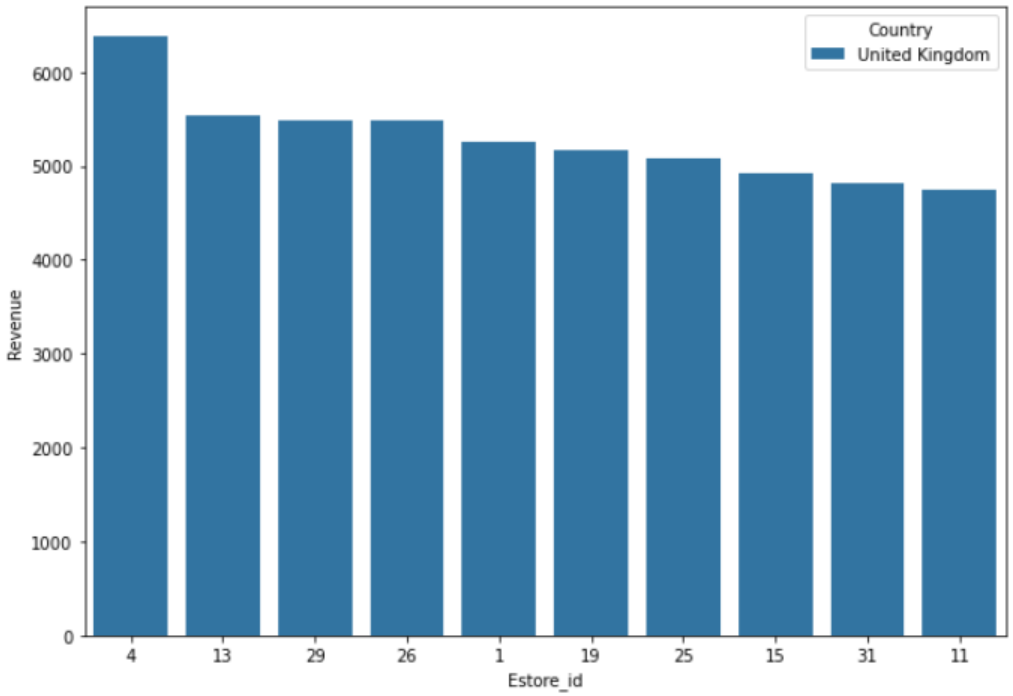
5. Top Countries contributing to revenue are UK, Germany, France, EIRE (Ireland), Netherlands.



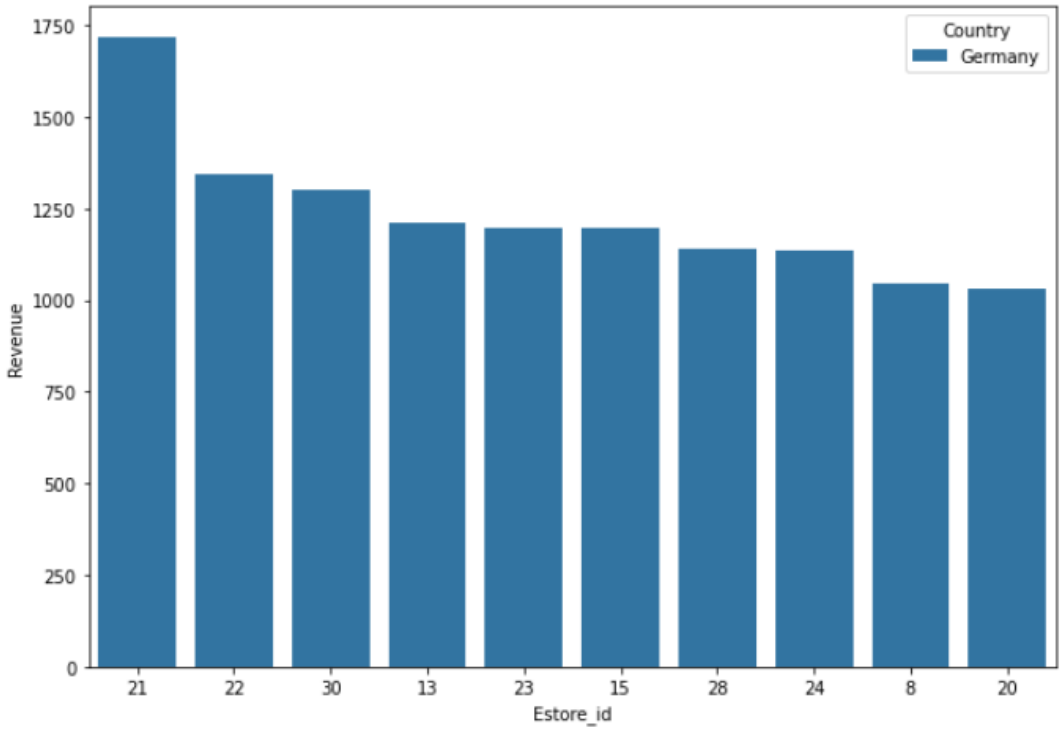
6. The top 4 countries as mentioned above contribute to more than 80% of revenue. Goal is to focus more on them in turn see why the other countries are not performing inspite of having the same number of transactions.



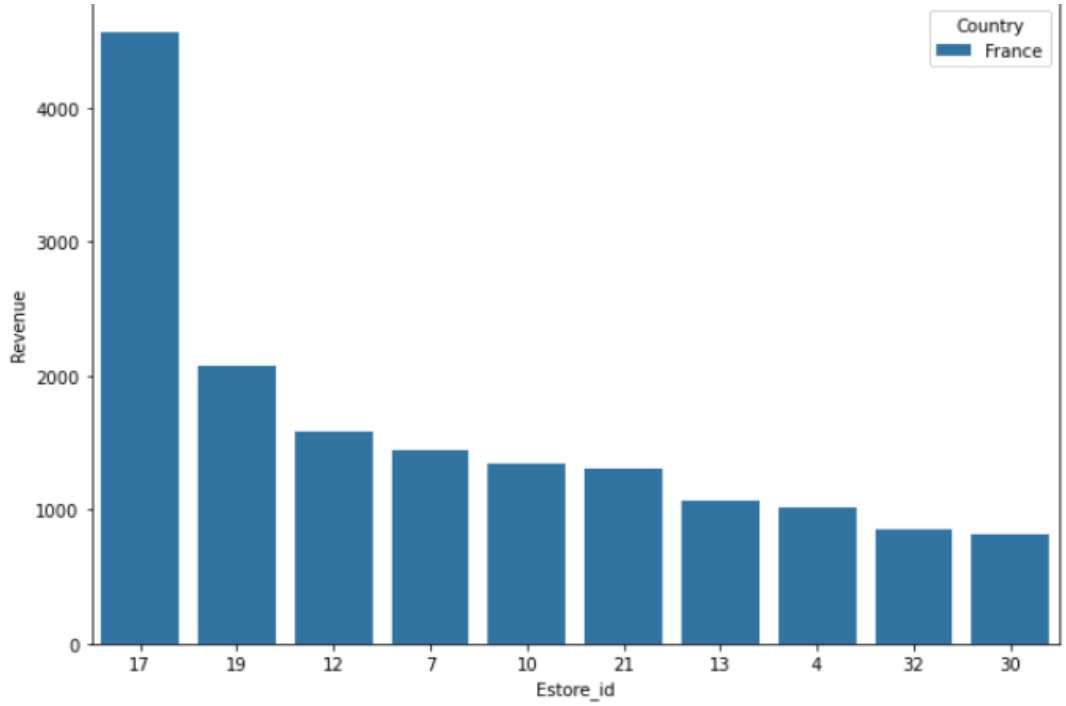
7. The top e-stores associated with each country are:
    a. United Kingdom



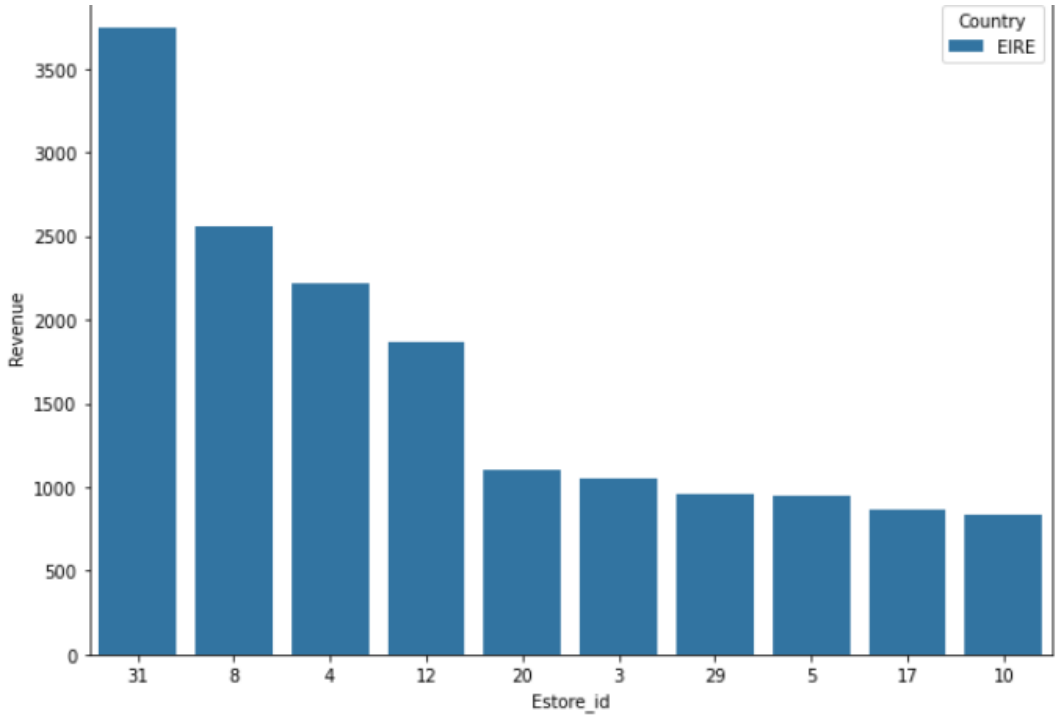
b. Germany



c. France



d. EIRE


Country
EIRE
Revenue
3500
3000
2500
2000
1500
1000
500
0
31
8
4
12
20
3
29
5
17
10
Estore_id

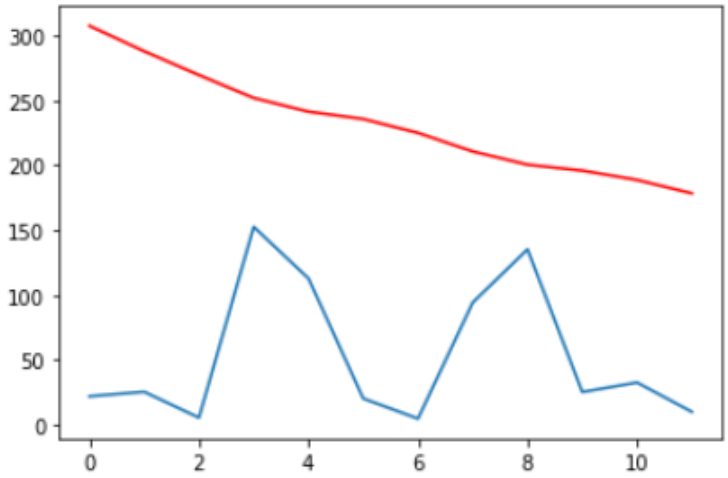
4. Forecasting

A Simple forecasting exercising has been performed on the data for the top three e-stores as mentioned above. For forecasting, ARIMA has been used with order 1,1,1 w.r.t order of autoregressive model, degree of differential and order of moving average respectively. For the ARIMA model, last 3 months has been taken for training and forecasts with the data being split on a 80-20 basis.

Forecasts for the top 3 e-stores are generated for both revenue and transactions. Forecasts for revenue fare worse than transactions. As seen below, the Test RMSE is better for transaction data than revenue. This can be improved by fixing data quality issues.

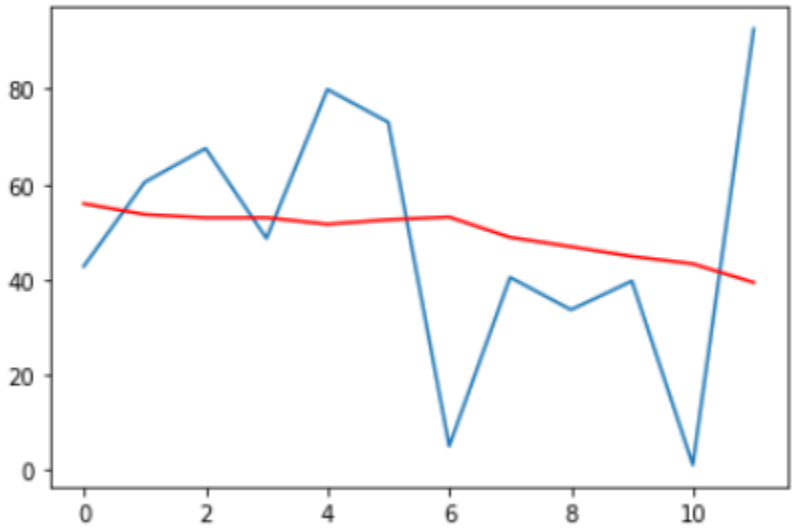
Revenue Forecast for e-store 1

```
Test RMSE: 191.86576345288714
```



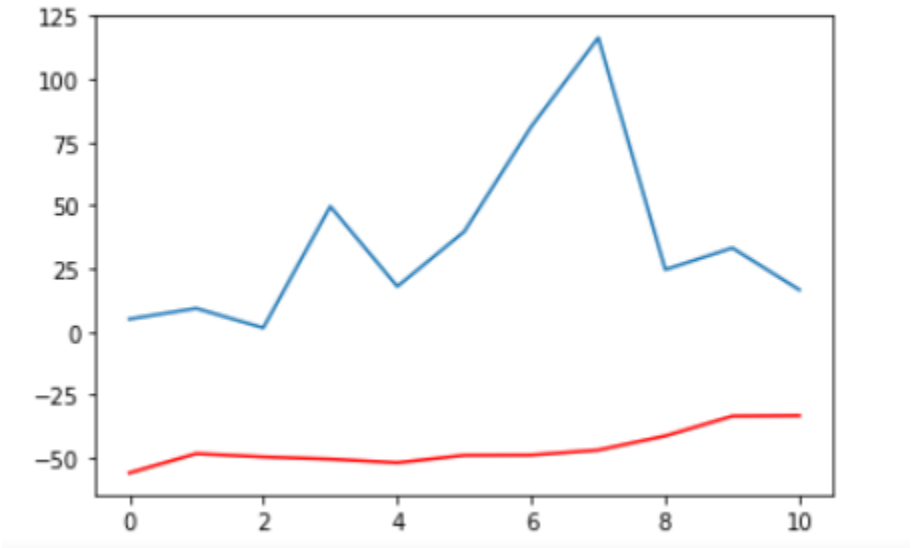
Revenue Forecast for e-store 31

```
Test RMSE: 27.183190614802747
```



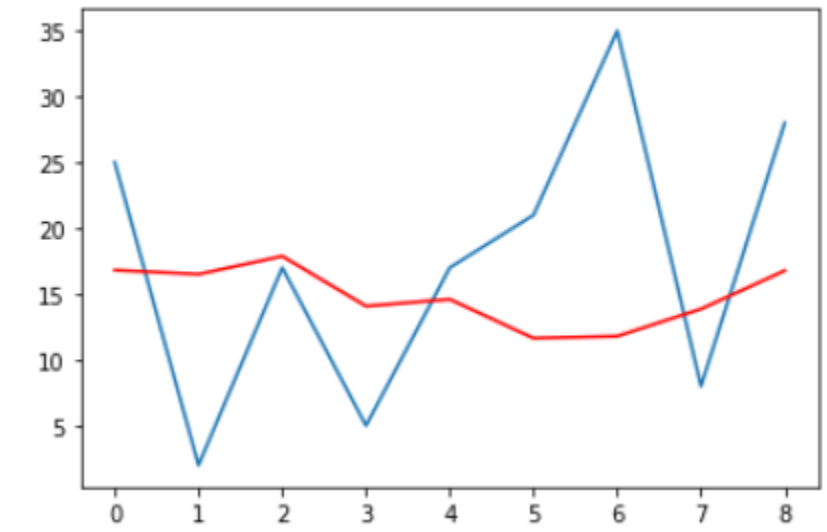
Revenue Forecast for e-store 4

```
Test RMSE: 89.06763816560242
```



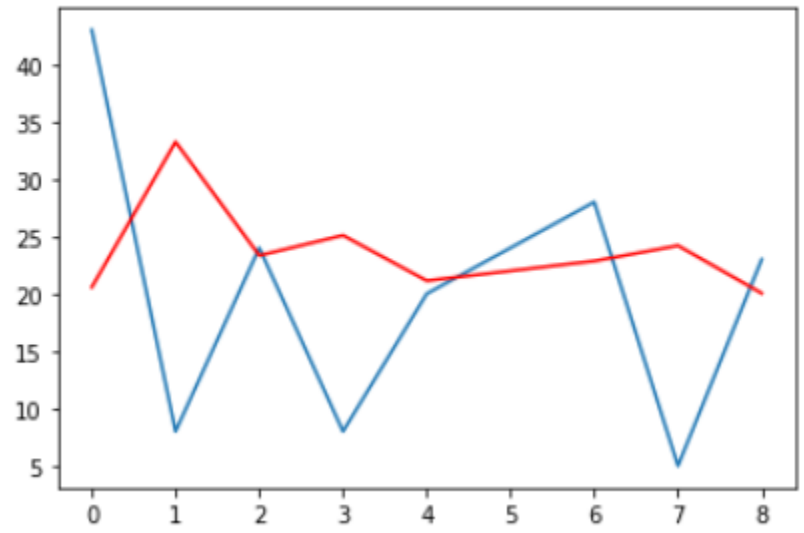
Transaction Forecast for e-store 1

```
Test RMSE: 11.317903033144953
```



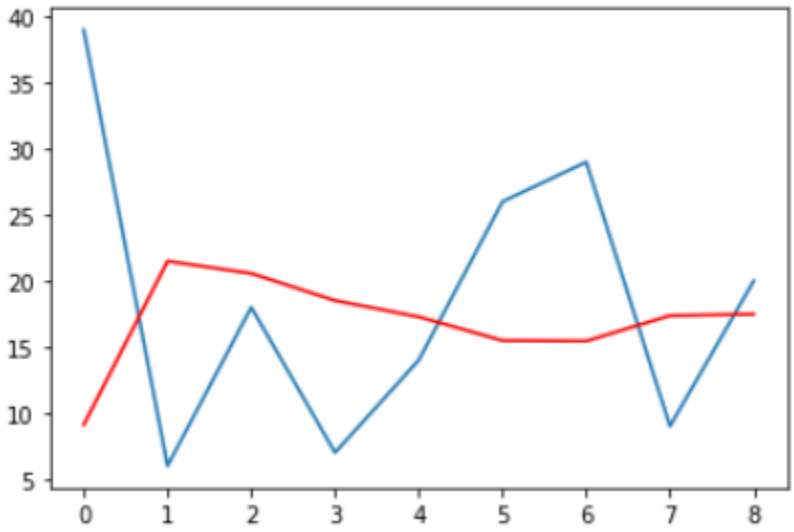
Transaction Forecast for e-store 31

```
Test RMSE: 14.305929602097565
```



Transaction Forecast for e-store 4

```
Test RMSE: 13.555098269756813
```



5. Recommendations and Future Scope

For improving the future scope of implementing the POC as a viable project, it can be improved in the following ways while working on the recommendations as well.

1. Forecasting works better for transactions as the number is quite constant as seen above in point 10 in EDA. Revenue prediction can be carried out with regression which could work better than revenue forecasting. It could also work with more data i.e., with improved quality field values.
2. Focus should be on generating more revenue from the top e-stores **1,31 and 4**, and countries like UK and Germany whilst also investigating the lower revenue generated from e-stores and countries like Italy, Belgium, and Norway even though they have same transaction counts
3. Similar case is with the top customers that contribute to the revenue of the top revenue generating e-stores while investigating the bottom customers. A study can be done about the customer satisfaction levels.
4. Improve quality of the transaction data.
    a. Less Null values.
    b. Avoid Negative values.
    c. Better descriptions.
5. With more time and better data, extend the forecast analysis to all the e-stores. Instead of doing ARIMA, a sequential order of forecast models for AR to Holt-Winters could have been followed.
6. API/Cloud for getting data instead of CSVs.
7. Descriptive Analytics can be hosted on any BI tool.
8. More customer data, so that we can perform a Customer Segmentation and Retention Analysis.
9. Market Basket Analysis (Association Rule Mining) based on description. Right now, there are many nulls and bigger text. It can help in GloboSales and e-stores to encourage more cross-buying of products.
10. Stocks are also to be focused on. More info is needed to understand StockCodes better.
11. More features based on transactions so that regression models can be used to predict revenue instead of a univariate forecast analysis.

Repository Link

All notebooks and the csv data has been saved in the github repo link below:

Assignment-LINK