SUPERSTORE DATASET

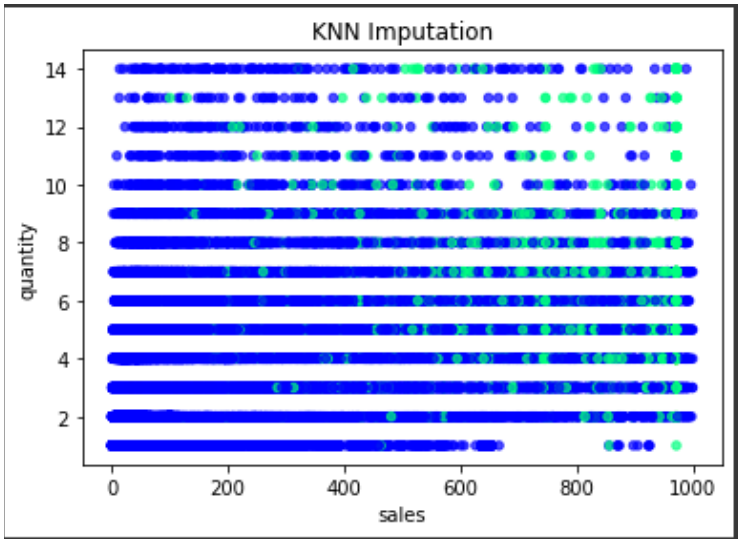
Introduction

The sample Dataset includes data for the Sales of multiple products sold by a super-store along with subsequent information related to geography, Product categories, and subcategories, sales, and profits, segmentation amongst the consumers, etc. This sample Dataset presents a common use case and also useful insights from the Sales data in order to improve the Marketing and Sales strategies.

Imputation for Null values

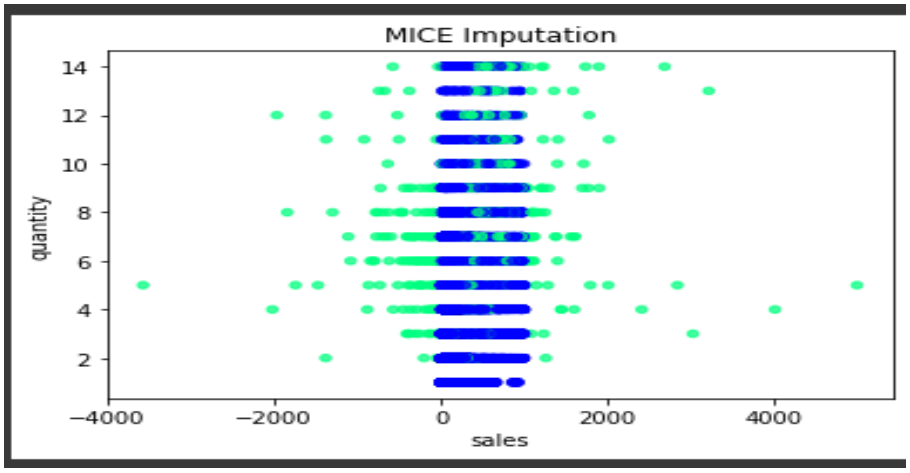
KNN Imputation

The idea in kNN methods is to identify 'k' samples in the dataset that are similar or close in space. Then we use these 'k' samples to estimate the value of the missing data points. Each sample's missing values are imputed using the mean value of the 'k'-neighbors found in the dataset.



MICE Imputation

MICE stands for **Multivariate Imputation By Chained Equations algorithm**, a technique by which we can effortlessly impute missing values in a dataset by looking at data from other columns and trying to estimate the best prediction for each missing value.

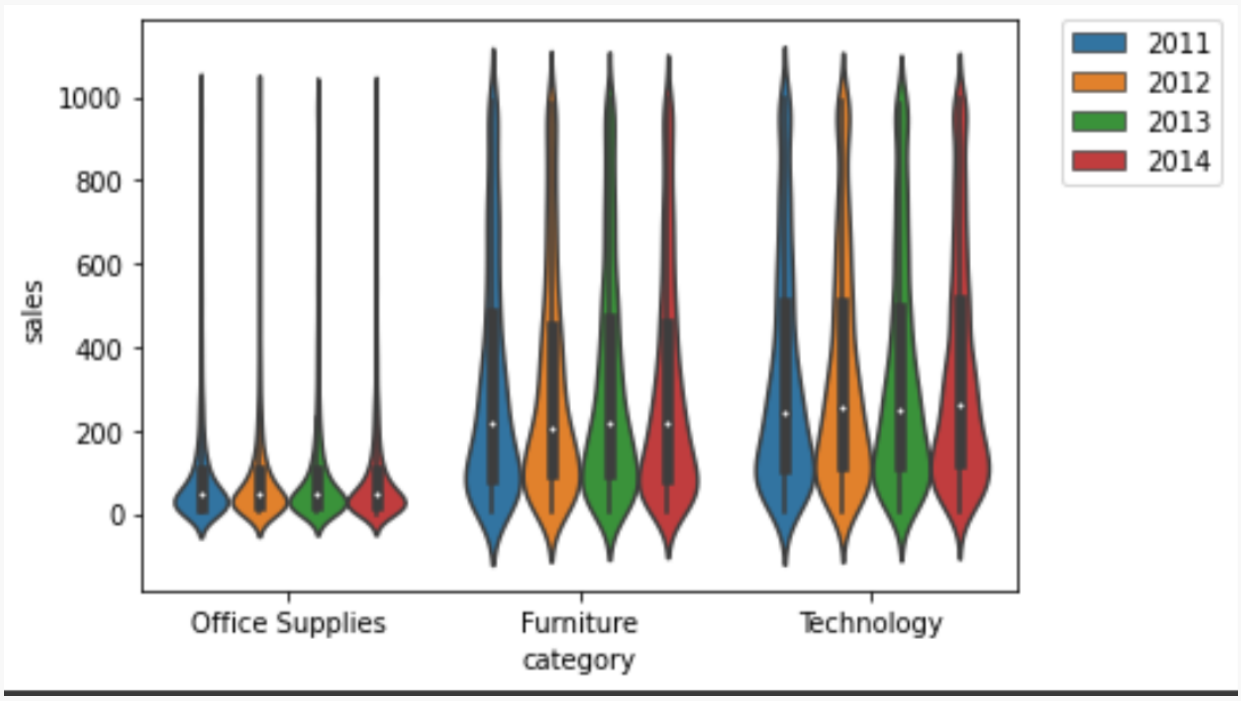


Here the green dots are the imputed values and the blue dots are the existing values. After getting the imputed values by two methods we clearly see that KNN imputation provides better fitting imputed values. As we can see, the green dots for MICE imputation are very scattered and certain values are very far from the already existing values(blue dots). Whereas in the KNN Imputation plot we see how well the green dots are well in place with the blue dots. Therefore we carry on with KNN imputation.

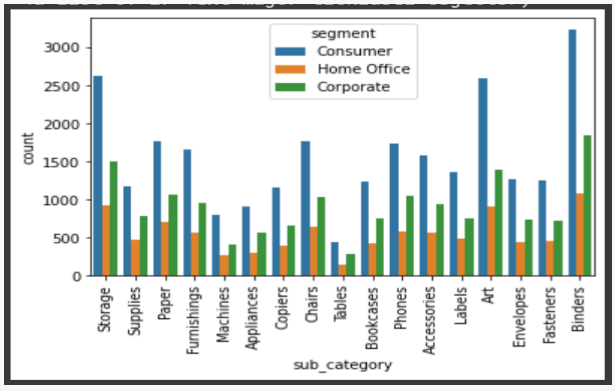
Analysis

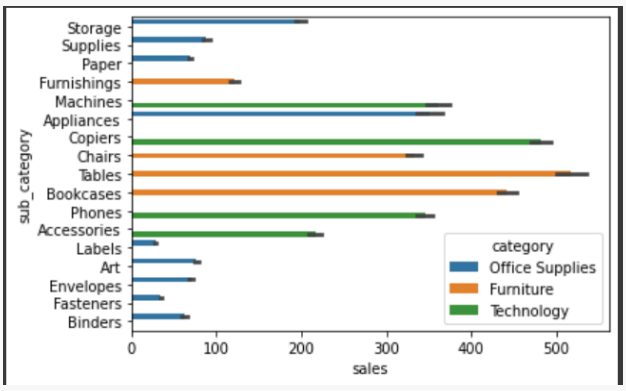
1. First we compare the sales year-wise and see there is not much difference in different years and then see the yearwise sales category - wise and see a certain difference in the trend of sales. **Technology** and furniture products perform considerably better than office supplies..

2. We then mark the trend for the frequency of products sold sub category wise divided by segments. Here we see that the **Binder** sub-category in the consumer segment has the highest frequency of orders whereas the **Table** sub category has the least frequency in all.



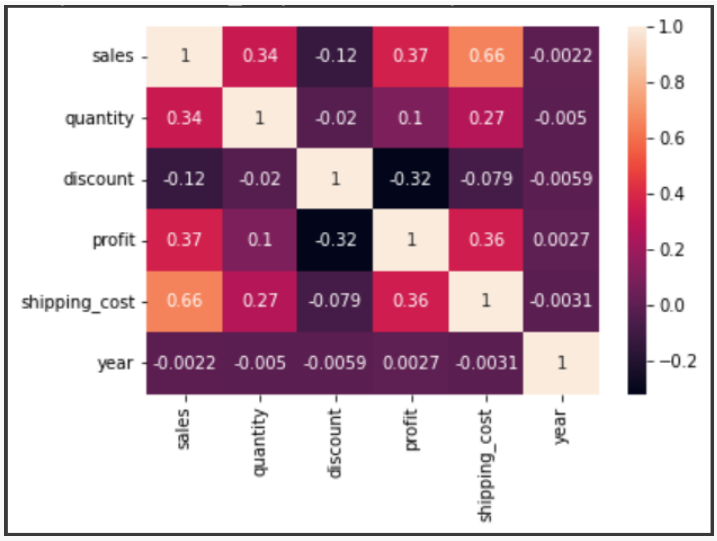
3. Now we have constructed a bar pot against sales and subcategory and have also marked the category to be clear. Here we see that even though **Table** had the least count of orders the sales for **Tables** under the Category Furniture is the Highest.





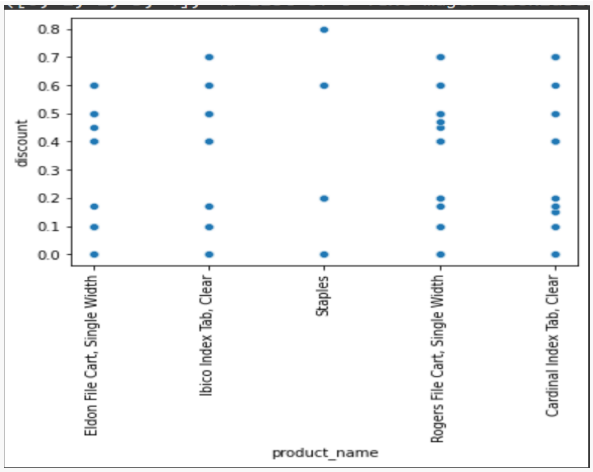
- Earlier when we did a year wise analysis the dataset clearly showed that the category **Technology** had the highest number of sales overall. But here when we divide the subcategories we see that the Sub category Table in the category Furniture has the highest sale .
- Office supplies have a very low rate of sales and it can be clearly analysed by all the graphs plotted above.
- Also we can see that the **Consumer** segment is more active when compared to other segments.It is followed by corporate and then home office.

4. Then we have a heatmap depicting no interesting correlations between features.



Product Analysis

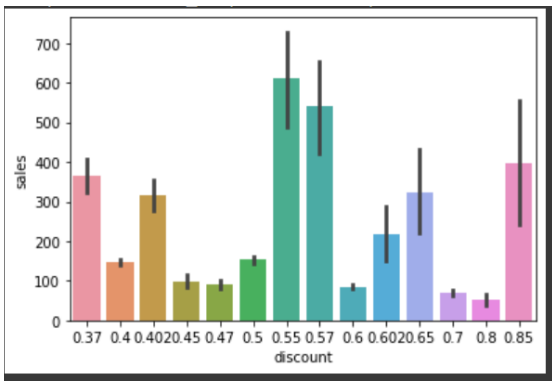
- First we find out the most ordered five products and find out the discount trend in the five products. We see that Staples have the highest discount. Whereas Rogers File Cart, Single Width has the most number of discounts.

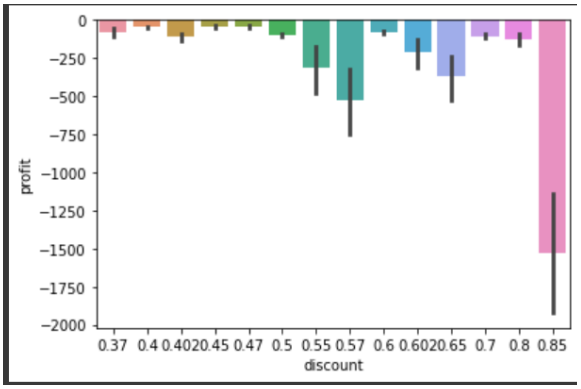


- While analysing discounts, we find that higher discounts are given out for a lesser number of products.

```
0.350    122
0.402    104
0.370     74
0.202     41
0.320     27
0.602     23
0.650     17
0.570     12
0.550     10
0.850      2
Name: discount, dtype: int64
```

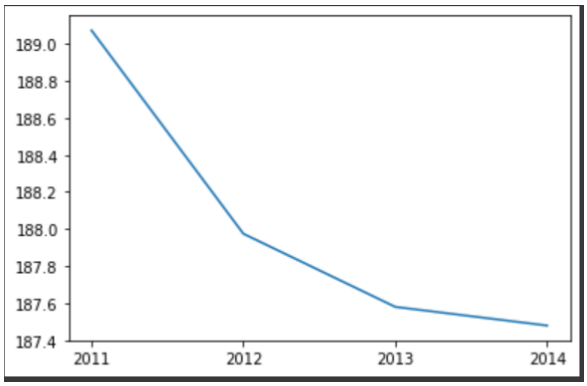
- Then we figure out the relation of sales with discounts offered to different products. We find that products in the band of 55-57% generate the most sales.
- We can also compare the discount with loss occurred due to the discount offered.
- The loss incurred in selling 2 products at a discount of 85 % is the most. Such details can be visualised from these plots and data.
- Also the major thing that the plot below depicts is offering discounts more than 35% for any product caused losses even if the sales were high enough. Therefore discounts in the mid-range can be given where discounts can be offered without causing losses to the store to attract customers.



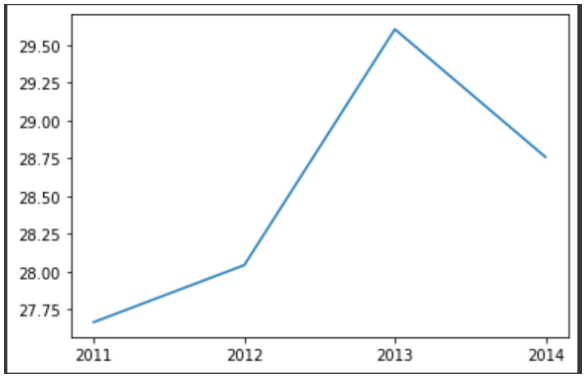


<u>OVERALL ANALYSIS</u>

- Average sales according to year is plotted on a line plot.

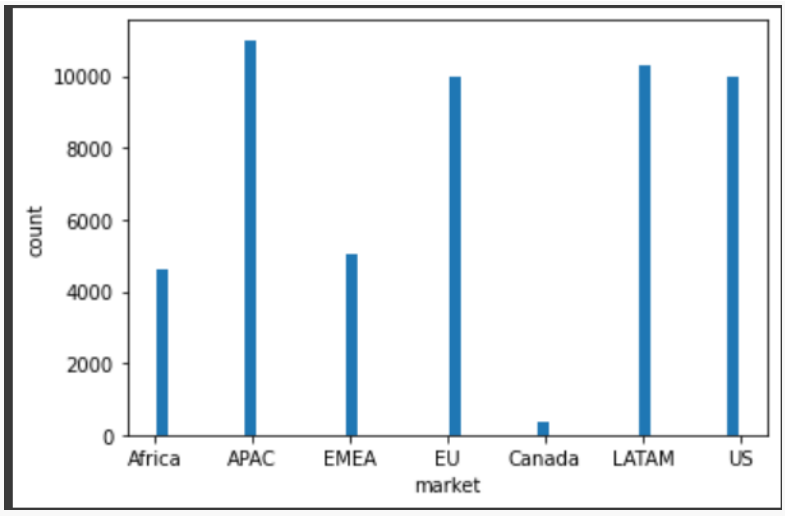


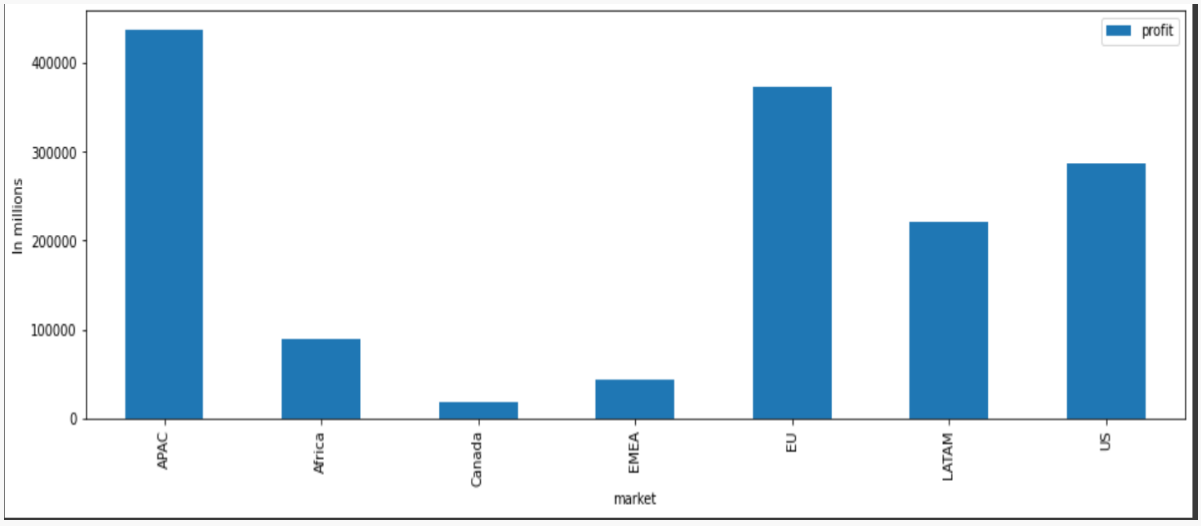
- Average profit according to year is plotted on a line plot.



We see here both average profit and average sales do not vary much in numbers due to the wide variety of products available and also the different segments the store has focused on. Also because of the discount trend and other similar factors the trend in the above features do not show a huge variation.

- Then finally we analyse the profit and frequency of orders market wise. Here we see how LATAM being the second highest on count still does not give high profit





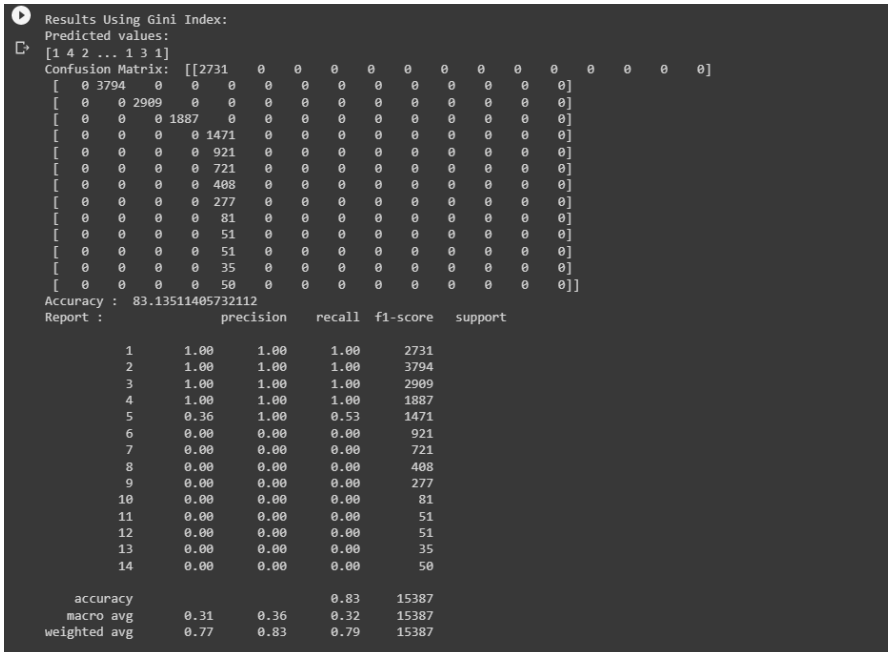
MODELLING
DECISION TREE USING GINI INDEX

One way of splitting a decision tree is via the Gini Index. The Entropy and Information Gain method focuses on purity and impurity in a node. The Gini Index or Impurity measures the probability for a random instance being misclassified when chosen randomly. The lower the Gini Index, the better the lower the likelihood of misclassification.
The formula for Gini Index

$$Gini = 1 - \sum_{i=1}^{j} P(i)^2$$

Where j represents the no. of classes in the target variable and P represents the ratio of the class at the ith node.The Gini index has a maximum impurity of 0.5 which is the worst we can get and maximum purity of 0 is the best we can get.

We have hence optimised the performance of the decision tree model using gini index. Therefore the accuracy we received was 83.13%.



```
Results Using Gini Index:
Predicted values:
[1 4 2 ... 1 3 1]
Confusion Matrix:  [[2731    0    0    0    0    0    0    0    0    0    0    0    0    0]
 [   0 3794    0    0    0    0    0    0    0    0    0    0    0    0]
 [   0    0 2909    0    0    0    0    0    0    0    0    0    0    0]
 [   0    0    0 1887    0    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0 1471    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0  921    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0  721    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0  408    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0  277    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0   81    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0   51    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0   51    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0   35    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0   50    0    0    0    0    0    0    0    0    0]]
Accuracy :  83.13511405732112
Report :               precision    recall  f1-score   support

           1       1.00      1.00      1.00      2731
           2       1.00      1.00      1.00      3794
           3       1.00      1.00      1.00      2909
           4       1.00      1.00      1.00      1887
           5       0.36      1.00      0.53      1471
           6       0.00      0.00      0.00       921
           7       0.00      0.00      0.00       721
           8       0.00      0.00      0.00       408
           9       0.00      0.00      0.00       277
          10       0.00      0.00      0.00        81
          11       0.00      0.00      0.00        51
          12       0.00      0.00      0.00        51
          13       0.00      0.00      0.00        35
          14       0.00      0.00      0.00        50

    accuracy                           0.83     15387
   macro avg       0.31      0.36      0.32     15387
weighted avg       0.77      0.83      0.79     15387
```

Therefore in the project above we have observed how different products, categories , sub-categories , segments cause differences in sales , profits and frequency of orders. However most of these are not correlated and hence do not follow a specific pattern but we can get an overview of how the store can maximise its sales/profits and attract more buyers.