“Sales Data Exploration, Reduction & Visualization”

Group No: 12
Branch: AIML
Batch: AIML 1
MEMBERS:

NAME	SAP ID
JAIMIN SHAH	60002190049
AAYUSH GANDHI	60002190006
DWAYNE GONSALVES	60002190040
DISHA KUNJADIA	60002190035
NEETI ADHIA	60002190069
YUKTI DOSHI	60002190125

TEAM NAME: NATASHA
UNIVERSITY NAME: DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

Table of Contents

Sr NO	Title	Page NO
1	**Background:**	
	1.1 Aim	3
	1.2 Technologies	3
	1.3 Software Architecture	4
2	**System:**	
	2.1 Requirements	4
	2.2 Design and Architecture	5
	2.3 Dataset and Variables	6
	2.4 Implementation	7-11
3	**Conclusions**	11
4	**Further development or research**	12
5	**References**	12
6	**Appendix**	13

Executive Summary

The project “Sales Data Exploration, Reduction & Visualization” was done to find the tie-in between sales and strategic marketing. By analyzing and exploring the data we aimed to help find an optimum way of increasing sales and at large help in targeting the perfect audience while marketing. For visualization and exploring data it first needs to be collected and cleaned. Then using different libraries and models the data was analyzed. We discovered many parameters on which the sales paradigm was balanced. Lastly, there was the prediction using the RFM model.

1. Background

1.1 Aim

Sales Data Exploration, Reduction & Visualisation.

Sales and marketing analytics are essential to unlock commercially relevant insights. These insights are necessary for increasing revenue and profitability and improving brand perception. A sales analysis report identifies the actual sales of a company over a period of time. The report shows if sales are increasing or declining. With an analysis, actual sales may be compared to expected sales.
An industry analysis allows businesses to estimate how much profit can be generated. Some questions to consider are:

- Size of the market
- How much the consumer spends
- How frequently the consumer spends
- The most preferred brand and most sold product

1.2 Technologies

We have implemented deep learning algorithms and RFM(random forest) for the sales data analysis and prediction in this project.

Deep learning is a class of machine learning algorithms that uses multiple layers to extract higher-level features from the raw dataset. It improves learning by examining on its own. Deep learning works with artificial neural networks, to imitate how humans think and learn.

1.3 Software Architecture

We used multiple libraries to make this project. Numpy, Pandas, Seaborn, Matplotlib, for the sales data analysis and prediction.

The **Pandas** module is mainly used to work with the tabular data. **The NumPy** module handles the numerical data. Pandas provide us with the powerful tools required for analyzing data, like DataFrame and Series. NumPy offers us the array.

Seaborn and Matplotlib libraries in **Python** are mainly **used** for making statistical graphics. **Seaborn is** a data visualization library so it was best used for bar plots, heat maps, and other visualizations. Visualization **is** the central part of **Seaborn and** it treats the entire dataset as a solitary unit. This was for the exploration of the cleaned data and prediction. The graphs showed us the pattern and behavior of the market.

2. System

2.1 Requirements

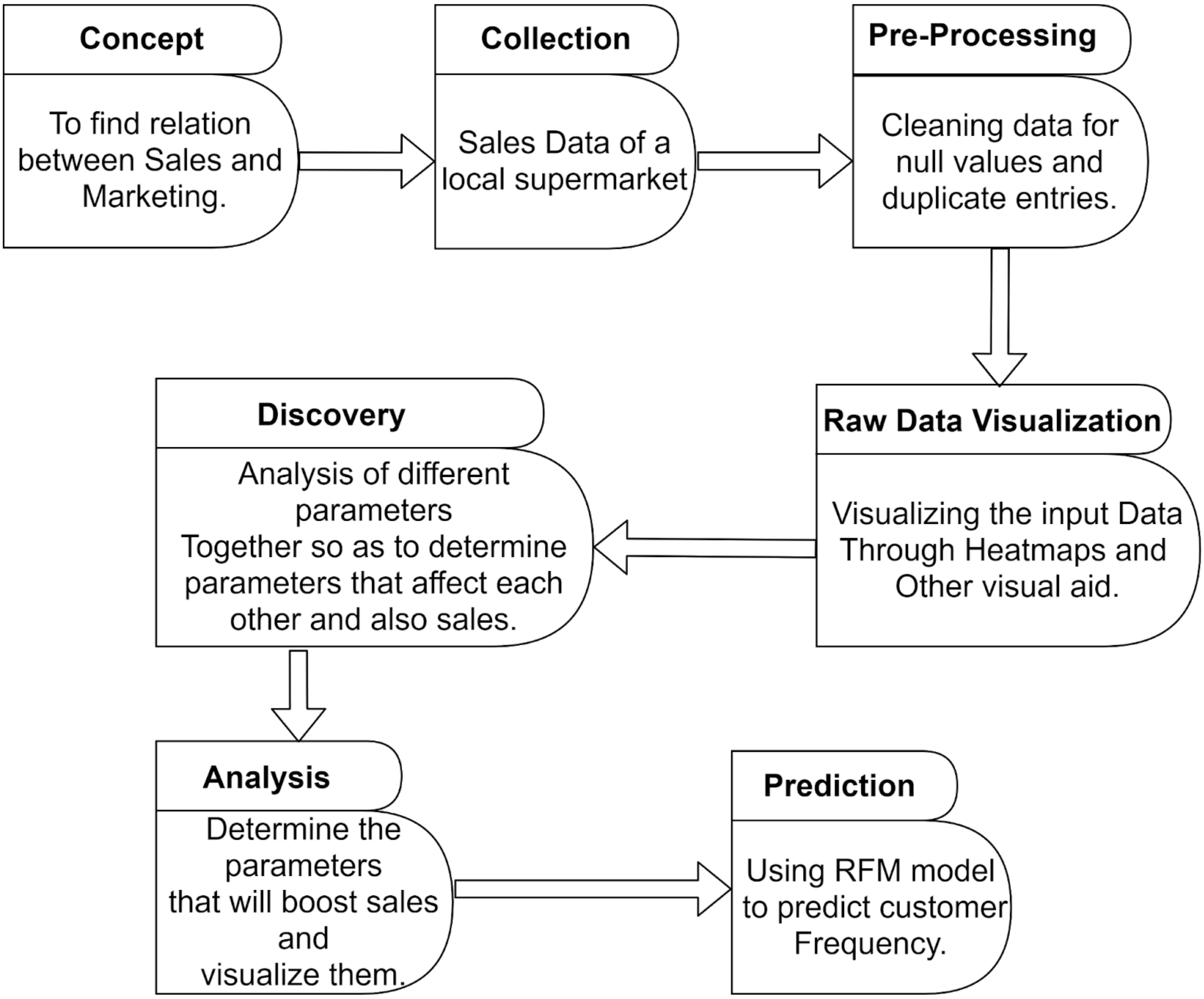
Python 3.8

Matplotlib 3.4.2

Pandas 1.2.5

Seaborn v0. 11.1

2.2 Design and Architecture


Concept
To find relation between Sales and Marketing.
Collection
Sales Data of a local supermarket
Pre-Processing
Cleaning data for null values and duplicate entries.
Raw Data Visualization
Visualizing the input Data Through Heatmaps and Other visual aid.
Discovery
Analysis of different parameters Together so as to determine parameters that affect each other and also sales.
Analysis
Determine the parameters that will boost sales and visualize them.
Prediction
Using RFM model to predict customer Frequency.

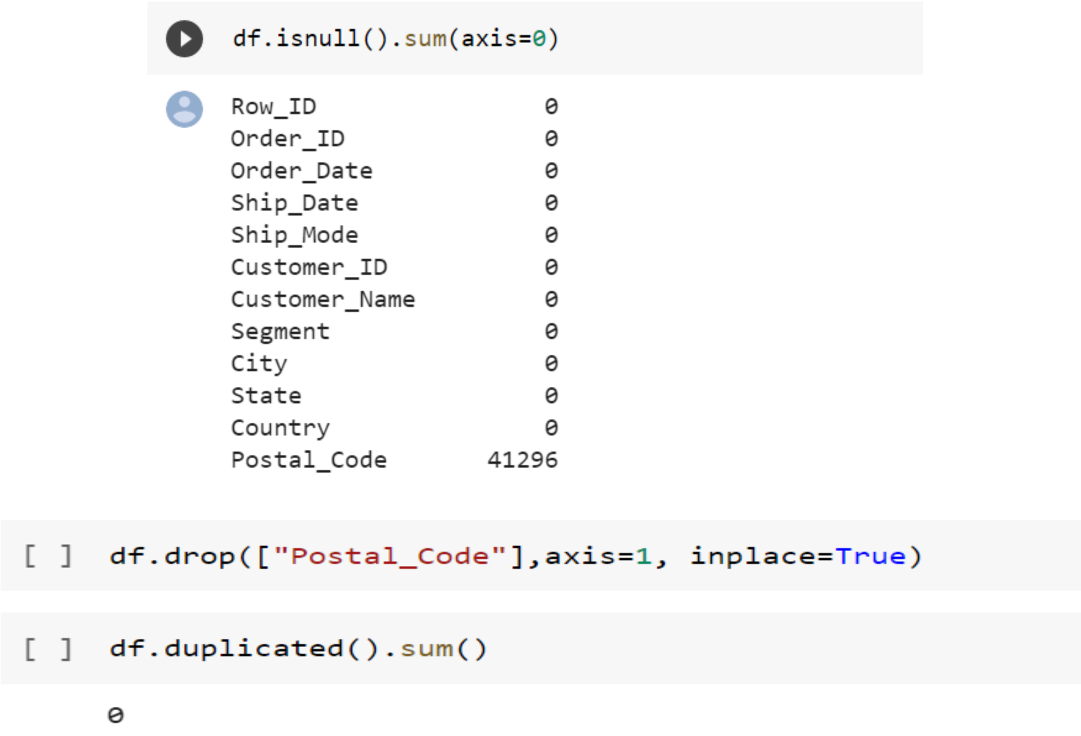
2.3 Dataset and Variables

Row_ID	Computer-generated sales slip invoice identification number
Order_ID	Computer-generated sales slip invoice identification number
Order_Date	Date of Purchase
Ship_Date	The date on which the product was delivered to the buyer.
Ship_Mode	Type of delivery opted by the customer
Customer_ID	Unique identification number of each customer
Customer_Name	Name of the customer
City	City from which customer ordered
State	state from which the customer ordered
Country	Country from which customer ordered
Product_ID	Unique identification number of the product
Category	Category to which the product belongs
Sub_Category	Subcategory to which the product belongs
Product_Name	Name of the product
Sales	Price of the product
Quantity	Quantity of product purchased
Discount	Discount available for the given Product
Profit	Profit earned by the sell
Shipping_Cost	The cost charged for the delivery of the items
Order_Priority	The products which are most selling or important to the customers should reach the earliest
year	Year of purchase
month	Month of purchase

2.4 Implementation

2.4.1 Preprocessing

The very first step for any analysis or visualization is to clean and organize the data. It also involves dropping unnecessary columns. Unnecessary columns are those with more than half of the values as null and have no relation to the analysis.



```
df.isnull().sum(axis=0)
```

```
Row_ID              0
Order_ID            0
Order_Date          0
Ship_Date           0
Ship_Mode           0
Customer_ID         0
Customer_Name       0
Segment             0
City                0
State               0
Country             0
Postal_Code     41296
```

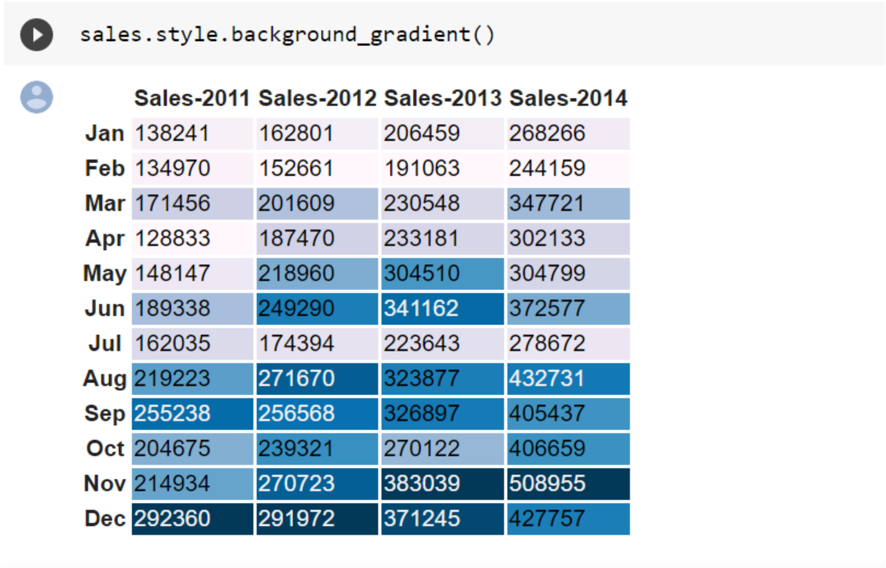
```
[ ]  df.drop(["Postal_Code"],axis=1, inplace=True)
```

```
[ ]  df.duplicated().sum()

     0
```

Postal code removal

2.4.2 Analysis

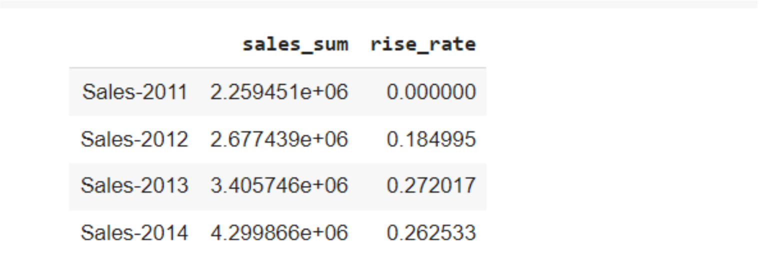
To analyze the data, it has to be grouped and rearranged.



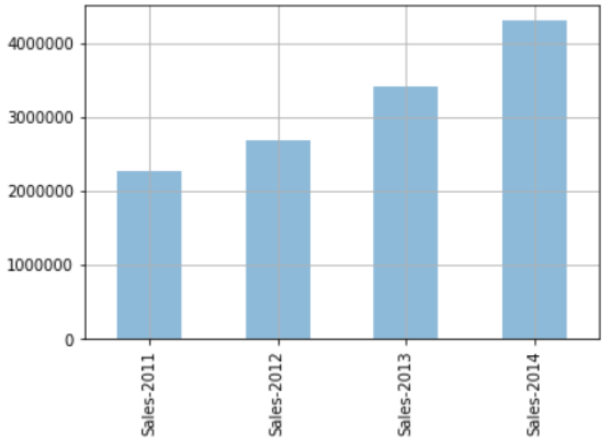
```
sales.style.background_gradient()
```

	Sales-2011	Sales-2012	Sales-2013	Sales-2014
Jan	138241	162801	206459	268266
Feb	134970	152661	191063	244159
Mar	171456	201609	230548	347721
Apr	128833	187470	233181	302133
May	148147	218960	304510	304799
Jun	189338	249290	341162	372577
Jul	162035	174394	223643	278672
Aug	219223	271670	323877	432731
Sep	255238	256568	326897	405437
Oct	204675	239321	270122	406659
Nov	214934	270723	383039	508955
Dec	292360	291972	371245	427757

From above, it is seen that there has been an accretion in the sales figure over the years. We don't need to have an amazing business acumen to conclude that sales performance has grown rapidly and development is really good.

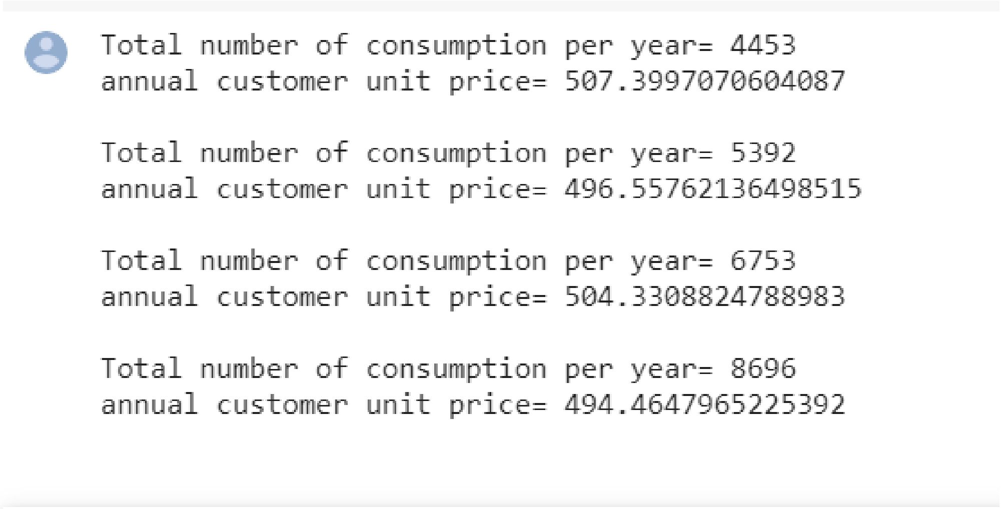


	sales_sum	rise_rate
Sales-2011	2.259451e+06	0.000000
Sales-2012	2.677439e+06	0.184995
Sales-2013	3.405746e+06	0.272017
Sales-2014	4.299866e+06	0.262533



It can be seen in the graph plotted that the sales growth rate in the next two years touched 26%. The sales in 2014 were nearly twice that in 2011. The development momentum is good and the operation is gradually stable.

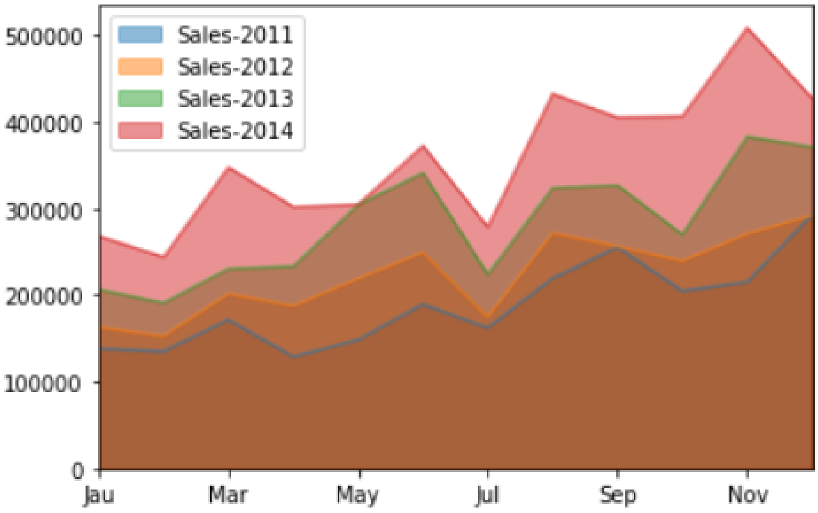
After understanding the overall sales of the supermarket, analyze the monthly sales each year to understand the sales in different months, find out whether there are low and peak seasons, and find out the key sales months in order to formulate business strategies and performance monthly and quarterly Indicator split.



```
Total number of consumption per year= 4453
annual customer unit price= 507.3997070604087

Total number of consumption per year= 5392
annual customer unit price= 496.55762136498515

Total number of consumption per year= 6753
annual customer unit price= 504.3308824788983

Total number of consumption per year= 8696
annual customer unit price= 494.4647965225392
```

From the above results, the number of consumption per year is on the rise, but the overall fluctuation range of the customer unit price is not very large, stable at about 500.

We have also analyzed that the number of new customers is decreasing year by year. It indicates that the company's old customers are well maintained, but the new customer acquisition rate is low.



It can be generally seen from the above picture, the grocery store's deals are occasional. In general, the start of the year is the slow time of the year and the subsequent half is the pinnacle season. Deals in June go high and have a fall in sales in July. Deals in July in the second 50% of the year were low. They are similar to the sales at the start of the year.

There is a peak in November and sales continue to remain high in December too. It is an indication that the rise at the end of the year is due to the festive season. But there was an unusual downfall in 2014, maybe due to the economic crash.

However for the peak season, strategies such as operation promotion should be maintained, and investment can be increased to increase overall sales. For the off-season months, new product development can be combined with product characteristics, and some promotional activities can be held to attract customers.

2.4.3 Prediction

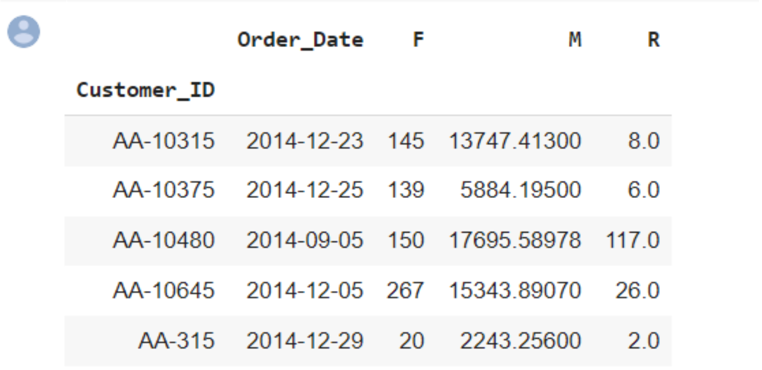
We used the RFM for the prediction of sales in the coming years. It was possible due to the Customer return prediction using RFM Model:

Recency, frequency, monetary value (RFM) is a marketing analysis tool that will identify a company's or an organization's best customers by using certain measures. The RFM model is based on three quantitative factors:

Recency: When did the customer make his last purchase

Frequency: How often a customer purchase

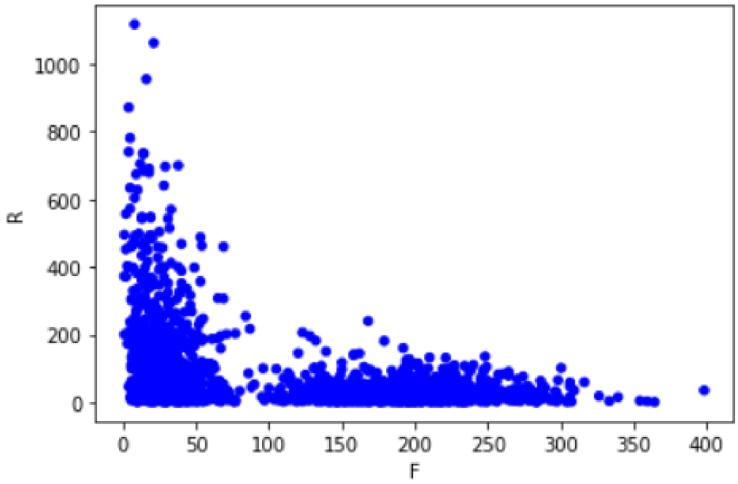
Monetary Value: How much money, the customer spends on the products



	Order_Date	F	M	R
Customer_ID				
AA-10315	2014-12-23	145	13747.41300	8.0
AA-10375	2014-12-25	139	5884.19500	6.0
AA-10480	2014-09-05	150	17695.58978	117.0
AA-10645	2014-12-05	267	15343.89070	26.0
AA-315	2014-12-29	20	2243.25600	2.0

Here also F is the frequency at which customers visit the store and buy products. M is the monetary value, where how much is spent and R is the recency, last when did the customer buy the products from that store.

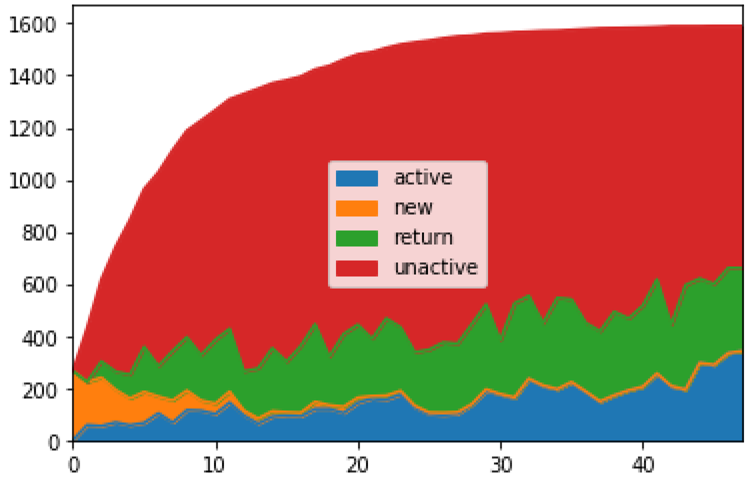
```
<matplotlib.axes._subplots.AxesSubplot at 0x2129e7d25c8>
```



Identify different customer groups through RFM, can measure customer value and customer profitability, can specify personalized communication and marketing services, provide strong support for more marketing decisions, and create greater benefits for the company.

RFM also helps us find a balance, let’s take an example. Suppose X is a regular customer and spends 20$ on average every time X visits. X visits the store 4-5 times a week. And let’s suppose Y is also a regular customer but not that frequent. Y visits the store only twice a week. But Y spends 250-300$ on average. So this can be segregated and dealt with fairly that Y’s monetary value is far more than X and thus the frequency will be secondary.

```
<matplotlib.axes._subplots.AxesSubplot at 0x2129e7a5488>
```



From the above results, we found the number of active customers, new customers, and returning customers and their fluctuations every year. This may be related to year-end and year-end promotions and requires more data to support it. At the same time, it can be found that the number of new customers is decreasing every year, indicating that this merchant has a low rate of acquiring new customers. If they can make a breakthrough in acquiring new customers, it will bring a lot of room for business growth.

3. Conclusion

A sales analysis hence is all about more than how much money your sales team generates for your business. You can learn about your customers and what makes them tick through market research analysis. You also get a glimpse into your sales team and their performance through sales analytics as well.

Thus, Regular sales analysis creates accountability, reveals insights about your customers, the traits of top performing sales reps, and more aspects that will improve your bottom line. The different types of sales analysis methods and a step by step strategy to perform sales analysis is also provided.

4. Further development or research

There is so much scope in the project topic sales data analysis and prediction. This project can be further extended to predict the sales of the products for upcoming years. It can further be made more precise with other advanced models like ARIMA. This can be proved useful for the owner of the store to decide on which products to continue and which products need changes in order to prove profitable. It is also possible to increase sales if we are able to build a model that detects the sentiment of the customer so as to know when they are most likely to purchase more goods. Predicting when the sales will be high for some particular products and a specific region and market is also possible with further research and improvements. With analysis of sales and its prediction, it is also possible to learn about human behavior and the cognitive parameters of humans at large. It can help progress the industry and also minimize wastage especially food and other products which can exhaust resources and are time-sensitive. Accenture Analytics, Gramener, Crayon Data are the top-notch companies that analyze and predict data. These companies earn a huge chunk of profit from analyzing data of various markets and customer behavior.

5. References

Books:

i)THE 80/20 & RFM ANALYSIS PLAYBOOK

ii)Data Mining Using RFM Analysis

Websites:

https://imotions.com/blog/analyze-heat-maps/

https://www.optimove.com/resources/learning-center/rfm-segmentation

https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp

6. Appendix

Dataset:
https://docs.google.com/spreadsheets/d/e/2PACX-1vRtm0116ui_JQd18yKuD_8FmcK4I32HharVgLaNVx8LfRwA2MK6IL8t8w8ljgE7LoVn-SL5iYX7R7o6/pub?gid=1469681410&single=true&output=csv

Code:
https://drive.google.com/file/d/1tgn2UjlDoxFHenRUhyqDw5V7MYr5si1j/view?usp=sharing