Exploratory Data Analysis Of Sales Data

PH 12 Rutumbara Chakor
Department Of Computer Science
MIT World Peace University
Pune(M.H.), India

PH 29 Shruti Agrawal
Department Of Computer Science
MIT World Peace University
Pune(M.H.), India

PH 31 Sejal Kadam
Department Of Computer Science MIT World Peace University Pune(M.H.), India

PH 33 Vedant Karle
Department Of Computer Science
MIT World Peace University
Pune(M.H.), India

PH 39 Shreeraj Patil
Department Of Computer Science
MIT World Peace University
Pune(M.H.), India

***Abstract*—Data analysis is a necessary process carried out at ABC stores to get information on product sales. In carrying out its operational activities, Store uses standard excel in managing product sales. However, this application has not been able to assist managers in producing the required reports. A way to overcome this issue is by analyzing sales data using various ML techniques and providing data visualization for better understanding. In addition, data mining techniques are used to find out statistics like the most sold product, the best month for sales, products that are sold together, etc. that can help the organization make data-driven decisions.**

The research will produce reports in the form of Interactive Dashboard Visualization that can be used to make better decisions.

Keywords—Exploratory data analysis.Sales Insights

I. INTRODUCTION

Data management and analysis are essential for every business. With proper knowledge and valid information, the efficiency of the sales process can be increased. The ABC store mainly deals with electronics and applications etc. Their accountants have used standard excel to manage and organize sales-related data. But this data management system has not been able to help in decision making, product analysis, and give valuable reports of yearly sales.

One way to achieve this can be performing exploratory data analysis on the data. Exploratory data analysis (EDA) analyzes and investigates data sets and summarizes their main characteristics, often implementing data visualization methods. Manipulating data to get answers becomes relatively easy to discover patterns, test a hypothesis, or check assumptions.

Some of our objectives are

1. To develop a good marketing strategy by analyzing the linkage between different products.
2. To increase sales of the Store.
3. To provide insights about the top-performing and underperforming products.
4. To generate revenue reports from sales activities.
5. To visualize sales data for better understanding.
6. To analyze market trends and predict product sales.

II. DATASET

We have taken our dataset from the Kaggle repository. It is a sample sales data of each month that includes information about customers' orders.
We have collected sales data for the years 2017 and 2019. The dataset includes columns such as Product Name, Order Date, City, State, Country, Sales, Qty, etc.

III . LITERATURE REVIEW

Ref no.	ML technique used	Dataset used	Results obtained	Research gap
[1] Alexander et al., 2017	Market Basket Analysis(to find attributes), Apriori Algorithm (association rules)	Synthetic dataset	The most significant value of multiplication of support and confidence took. E->D and E->C items with support=42.86% and confidence=75%	Can increase dataset size, predict/identify future sales, and use sales data visualization for better understanding
[2] B. Ida Seraphim et al, 2018	Market Basket Analysis, AIS, Apriori, Apriori Dynamic, FP Growth, Multi-Level Association	Real dataset	Apriori Dynamic was found to be the most suitable as the no. of scans was reduced to 2	The system can predict how much money does a person is willing to spend on a particular item
[3] Dr. Zainab Pirani et al, 2017	linear regression, logistic regression, Market basket analysis	Synthetic dataset	Identified product trends and patterns	The system can add specific visualizations to display product sales and daily turnover
[4] Nargish Guptaet al, 2014	Market Basket Analysis	Real dataset	A daily report of product sales , customer buying patterns	The system can add regression methods and clustering.
[5] Wenhui Shan, 2020	Data Analysis and Forecasting under Big Data	Synthetic Dataset	organize and summarize data information in the early stage to establish a data model	Can integrate risk assessment into production and sales, can also combine sophisticated management methods to scientifically estimate sales risk issues and formulate countermeasures to improve the stability of The sales environment.
[6] Natalya et al, 2019	data representation, data analysis, data interpretation, correlation analysis, process automation, retail sales, management, machine-learning	Real Dataset	The data described by an a large number of features can be represented in a form suitable for the machine-learning methods application.	As a method of data pre-processing one of the popular approaches are correlation analysis was investigated. Correlation the analysis allowed to feature space dimension reduction.
[7] Ahmed Roshdy, Nada Sharaf, Madeleine Saad and Slim Abdennadher 2018	Visualization, Big Data	Real Dataset	web-based interface for visualizing data. The approaches aimed at a generic web-based platform having a flexible engine for data Representation. The output of the visualization could also, it can be used as a plugin that would act as an element in a bigger Project.	interaction ideas such as drag and drop features in which the data can be added and removed based on simple dragging from or to the chart diagrams displayed should be added.

[8] Jian Guan, Jun Cai, Jikun Liu 2013	user insight, opportunity-driven innovation, user research, user motivations	Real Dataset	The researchers' understanding of "Insight," as well as the process and methods used during insight discovering and development are crucial to achieving a guiding the purpose of opportunity-oriented user research for product and service innovation	The use of the insight discovering and development research method resulted in opportunity directions, which are helpful for future design applications, and several new products concepts are being developed by the client.
[9] Vitaliy Buyar, Amal Abdel-Raouf, 2019	Big Data, Neural Networking	Real dataset	Model is trained and validated. Finally, the results are evaluated based on the mean. absolute error and mean fundamental percent error metrics, which are used to determine the accuracy and show the effectiveness of the model.	Aggregate, raw sales data and train a model using Convolutional Neural Network

IV. METHODOLOGY

A. Linear Regression:

It is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. [10] There is only one independent variable(X) in simple linear regression, and based on that dependent variable, (y) is predicted.[11] Multiple Linear Regression — In multiple linear regression, there are numerous independent variables(X), and based on that, the dependent variable (y) is predicted.[11]

We used simple linear regression to predict the quantities ordered for a special price of the product. In our case, quantities ordered are the dependent variable, and the price is the independent variable. We input the product's price and get the quantities that might be ordered based on prediction. As price is a crucial factor for the sales of a product, this model of prediction is beneficial to making business decisions.

B. Random Forest:

It is an ensemble technique capable of performing regression and classification tasks using multiple decision trees and a method known as Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine numerous decision trees to determine the final output rather than rely on individual decision trees.
Random Forest has multiple decision trees as base learning models. [12]

We have used random forest and linear regression to predict the sales of some particular months, and we have compared them against the actual sales data.

In random forest regression, some inputs are given to train the model, and the model for the remaining inputs generates output.

On comparing random forest regression and linear regression, it was found that spontaneous forest regression is more accurate than linear regression for predicting sales.

V. TABLEAU VISUALISATION

Tableau:

Data visualization software is used to generate insights to understand the data better. There are many graphs that we can choose from and also create exciting dashboards. Such dashboards are helpful for the company to make business decisions.

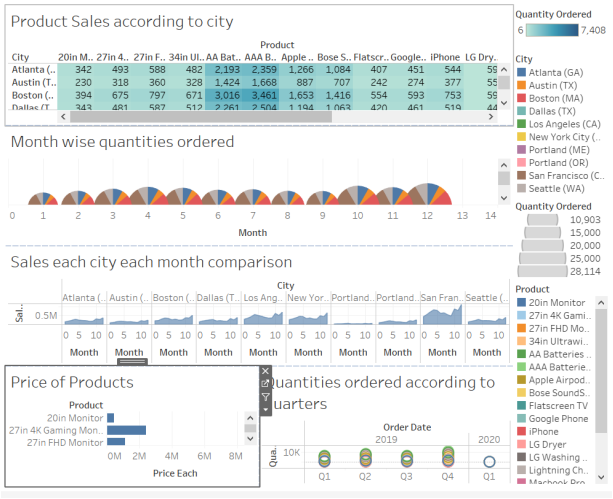

Figure 1.1 Tableau Dashboard

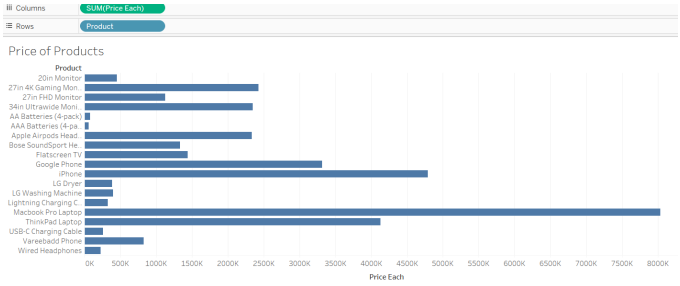

Figure 1.2 Indicating the price of products.

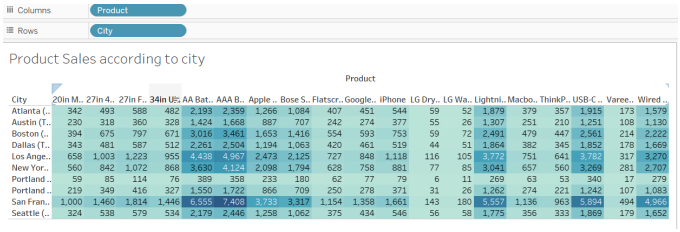

Figure 1.3 shows the product sales according to cities.

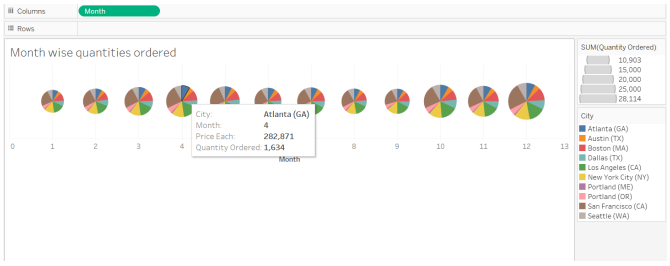

Figure 1.4 represents the month-wise quantities ordered for a particular city.

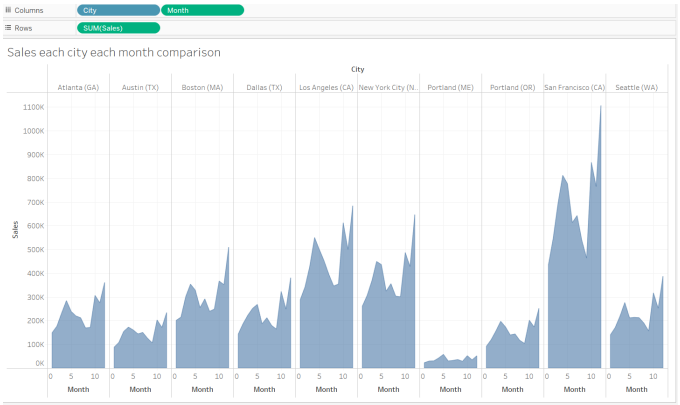

Figure 1.5 demonstrates the range of sales for each city and each month.

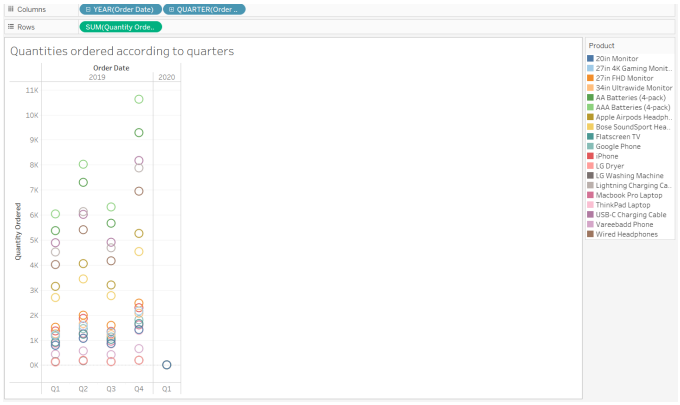

Figure 1.6 shows the number of products that are ordered in each quarter.

VI. RESULTS AND DISCUSSION

We had grouped sales by the month to see the total quantity ordered and total sales for respective months and used visualization to represent it. For our dataset, we could see that the best month for sales was December, with a revenue of around 4.5 million.

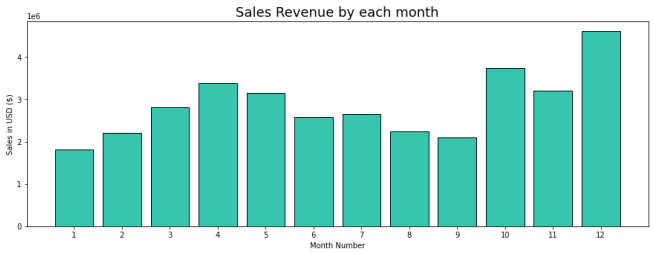

Figure 2.1 Sales Revenue By Month

San Francisco had the highest sales of about 8.2 million, followed by Los Angeles with 5.4 million revenue.

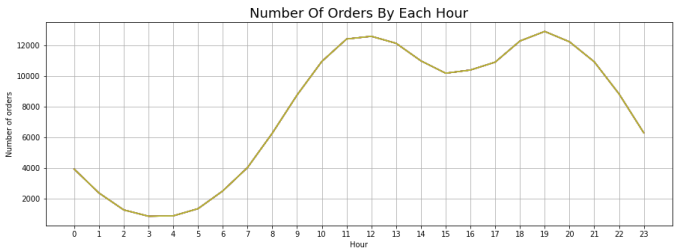

Figure 2.2 Order Number By Each Hour

Our analysis found out that more than 12,000 orders were placed at around 7 pm. Hence it would be the best time to advertise the products.

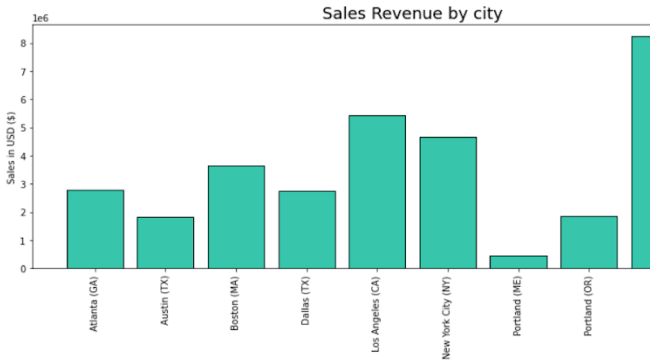

Fig 2.3 City wise sales

	Product Pair	Quantity Ordered
0	(iPhone, Lightning Charging Cable)	1005
1	(Google Phone, USB-C Charging Cable)	987
2	(iPhone, Wired Headphones)	447
3	(Google Phone, Wired Headphones)	414
4	(Vareebadd Phone, USB-C Charging Cable)	361
5	(iPhone, Apple Airpods Headphones)	360
6	(Google Phone, Bose SoundSport Headphones)	220
7	(USB-C Charging Cable, Wired Headphones)	160
8	(Vareebadd Phone, Wired Headphones)	143
9	(Lightning Charging Cable, Wired Headphones)	92

Figure 2.4 Products Sold Together

We grouped the sales data with the same order ID to find an association between the products. We got the result that iPhone and lightning charging cable was ordered 1005 times, followed by Google phone and USB-C charging cable, which was called 987 times.

From our analysis, we conclude that the most sold product was AAA batteries(4-pack), the reason being the product's low price. It is visible from the graph that the cheapest products were sold the most.

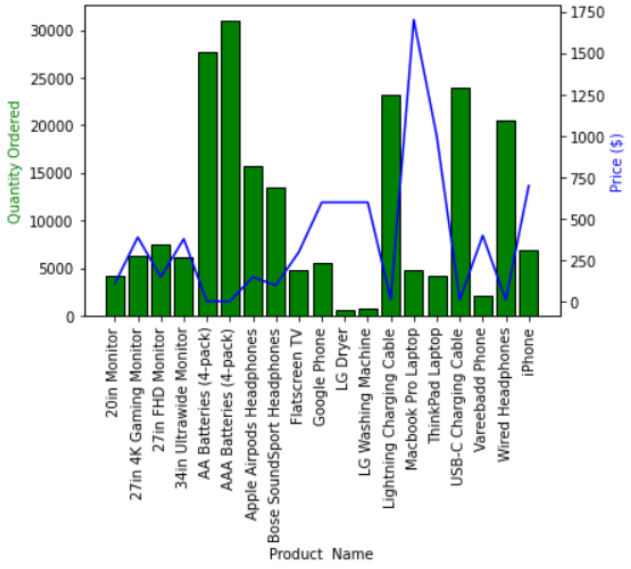

Figure 2.5 Comparison of price each and product sold

We used a linear regression model to determine the quantity ordered for a product according to the price set. The image shows a demonstration of the same :

```
[35] reg.predict([[500]])

     /usr/local/lib/python3.7/dist-packages/sklearn/ba
       "X does not have valid feature names, but"
     array([9383.97857081])
```

Figure 2.6

We have found out the city in which a particular product is sold the most. It can be seen that USB-C Charging Cable , Bose SoundSport Headphones, AAA batteries, and iPhones are most sold in San Francisco followed by Los Angeles and New York City.

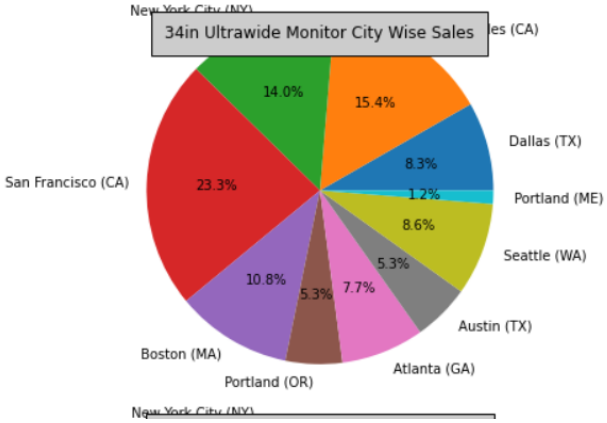

Figure 2.7.1 Reference 1

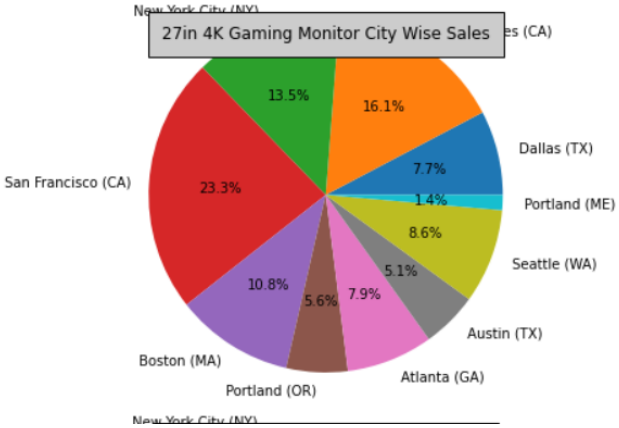

Figure 2.7.2 Reference 2

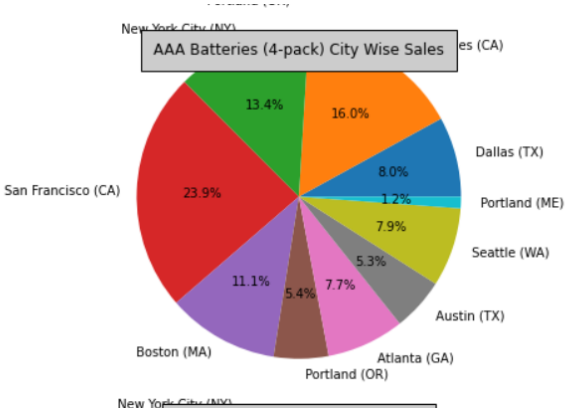

Figure 2.7.3 Reference 3

Sales Prediction :

We predicted sales for 2017 and 2019 using Random Forest and Linear Regression Models. We trained the dataset with six months of data and predicted the sales for another six months.

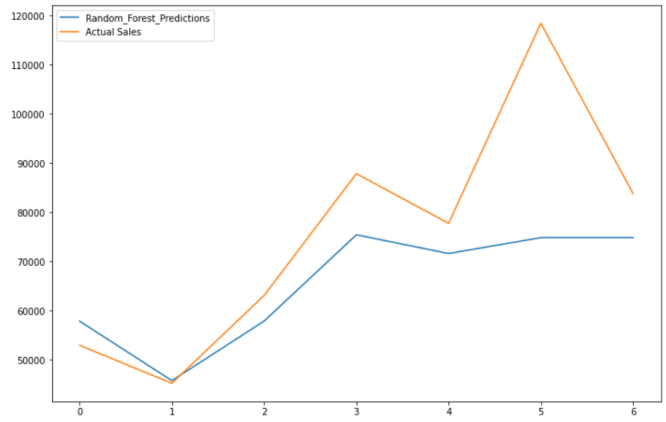

Fig 2.8.1 2017 Sales Prediction using random forest

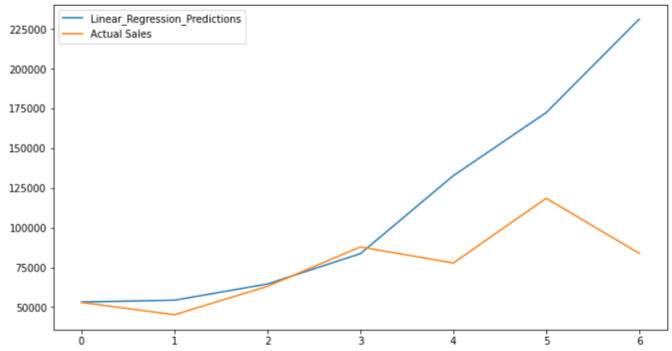

Fig 2.8.2 2017 sales prediction using linear regression

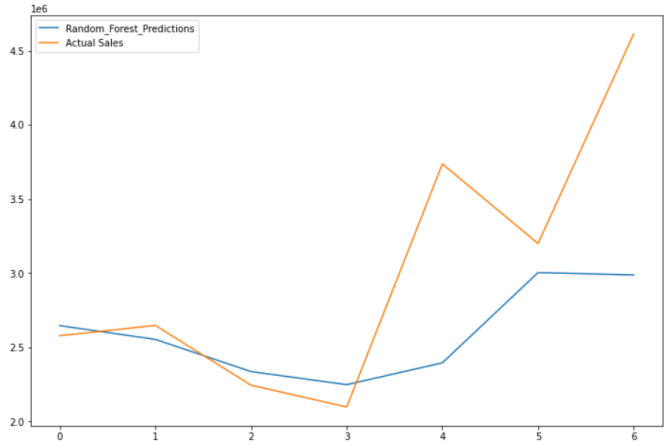

Fig 2.8.3 2019 Sales Prediction using random forest

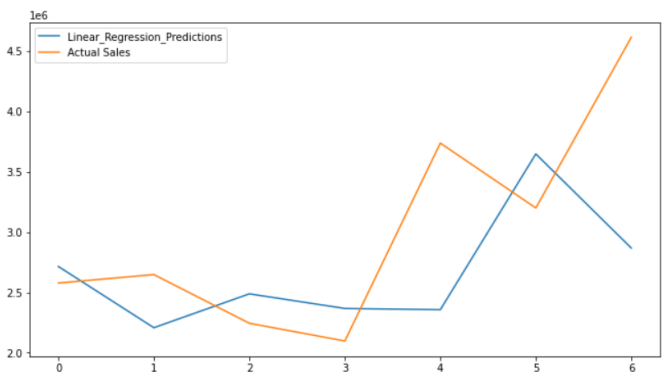

Fig 2.8.4 2019 Sales Prediction using linear regression

As evident from the figure, linear regression and random forest both performed well in the sales prediction for 2017, and random forest performed better in the sales prediction for 2019.

VII. FUTURE WORK

In the future, we plan to track which customer is buying what, when they are buying it, and at what price. We can determine how much a person belonging to a particular segment is willing to invest in our products. Data after analysis can be used in stores. Customer coupons target their buying habits and decide when to put items on sale or sell them at a total price.

VIII. ACKNOWLEDGEMENT :

We would like to thank Prof.Suja Panikar, who helped us in our project and cleared our doubts from time to time.

IX. CONCLUSION :

By analyzing the company's sales data, this study makes it helpful in making data-driven decisions by the officials to increase their sales in the upcoming months. To summarize, we found out that :

- The most sold product of the company is AAA batteries.
- Two products bought together most frequently are iPhone and lightning charging cable.
- Most sales for the company come from San Francisco, followed by Los Angeles and New York City.
- The best time for sales is the month of December.
- Random Forest Regression performed better than Linear Regression for predicting future sales.

This application can perform the data mining process based on existing sales data and thus provide the required reports to the manager.

X. REFERENCES :

[1] Alexander Setiawan; Gregorius Satia Budhi; Djoni Haryadi Setiabudi; Ricky Djunaidy," Data Mining Applications for Sales Information System Using Market Basket Analysis on Stationery Company".Pages:1-4.Available at- https://ieeexplore.ieee.org/document/8262592 Published in: 2017 International Conference on Soft Computing, Intelligent System and Information Technology (ICSIIT)

[2] B. Ida Seraphim, Lavi Samuel Rao, Shiwani Joshi, "Survey on Customer-Centric Sales Analysis and Prediction." Pages:1-6. Available at-https://ieeexplore.ieee.org/document/9034342 Published in 2018 3rd International Conference on Inventive Computation Technologies (ICICT)

[3] Dr. Zainab Pirani, Anuja Marewar, Zainab Bhavnagarwala, Madhuri Kamble, "Analysis and optimization of online sales of products."Pages:1-5.Available at - https://ieeexplore.ieee.org/document/8276165 Published in 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)

[4] Nargish Gupta, Madan Lal Yadav, An Implementation and Analysis of DSR Using Market Basket Analysis to Improve the Sales of Business.Pages:1-5.Available at: https://ieeexplore.ieee.org/document/6949249 They were published in the 2014 5th International Conference - Confluence The Next Generation Information Technology Summit (Confluence).

[5] Wenhui Shan, Research on Refined Sales Management, Data Analysis and Forecasting under Big Data. Pages:1-4. Available at: https://ieeexplore.ieee.org/document/9360971 Published in 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)

[6] Natalya V. Razmochaeva, Dmitry M. Klionskiy, Data Presentation and Application of Machine Learning Methods for Automating Retail Sales Management Processes. Pages:1-5. Available at: https://ieeexplore.ieee.org/document/8657077 Published in 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)

[7] Jian Guan; Jun Cai; Jikun Liu," Discovering and developing the "Insight" by opportunity-oriented user research."Pages:1-4.Available at : https://ieeexplore.ieee.org/document/6981231 Published in:2013 IEEE Tsinghua International Design Management Symposium

[8] Ahmed Roshdy; Nada Sharaf; Madeleine Saad; Slim Abdennadher, "Generic Data Visualization Platform."Pages:1-5.Available at:https://ieeexplore.ieee.org/document/8564138 Published in:2018 22nd International Conference Information Visualization (IV)

[9] Vitaliy Buyar, Amal Abdel-Raouf, "A Convolutional Neural Network-based Model for Sales Prediction."Pages:1-5.Available at:https://dl.acm.org/doi/10.1145/3388218.3388228 Published in: AIRC '19: Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control

[10] IBM, "Generate predictions using an easily interpreted mathematical formula," [Online]. Available: https://www.ibm.com/in-en/analytics/learn/linear-regression

[11]Y. Pandya, "Linear Regression — explained in simple terms!! | by Yagnik Pandya | Analytics Vidhya," 24 January 2021. [Online]. Available: https://medium.com/analytics-vidhya/linear-regression-explained-i n-simple-terms-Yagnik-8f9eccb680ec. [Accessed 3 February 2022].

[12]A. Dutta, "Random Forest Regression in Python," 18 January 2022. [Online]. Available: https://www.geeksforgeeks.org/random-forest-regression-in-pytho n/.