**Internship Report**

**Submitted to the Faculty**

**Of**

**St. Joseph’s College**

**(Autonomous)**

**Bengaluru Central University**



**By**

**Muhammed Hisham T, 21BDA43**

**In partial fulfillment of the**

**Requirements for the**

**Degree of**

**Master Of Science**

**In**

**Big Data Analytics**

**April 2023**

**I WOULD LIKE TO EXTEND MY SINCERE GRATITUDE TO THE**

**ALMIGHTY, MY PARENTS, AND MY PROJECT MENTORS.**



**DECLARATION OF THE CANDIDATE**

I hereby declare that this work entitled, "Guvi Sales Forecasting" has been originally carried out by me, Muhammed Hisham T under the guidance and supervision of Mr. Praveen Kumar J. This work has not been submitted elsewhere for the award of any other degree or diploma certificate.

June 2023 Muhammed Hisham T (21BDA43)

Bengaluru Department Of Advanced Computing

St. Joseph’s University

Bangalore

# ACKNOWLEDGEMENTS

This internship research work was carried out as part of my fourth semester M.Sc. Big Data Analytics coursework couldn't have been completed without the support of the following people.

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Mr. Praveen Kumar, for generously dedicating his valuable time to assist me in completing this project. Without his extensive knowledge and guidance, this work would not have reached the level of achievement it has attained.

I would like to express my gratitude to all the faculty members of St. Joseph’s University for their guidance throughout my postgraduate studies.

Lastly, I would like to sincerely express my deepest appreciation to my family and friends for their unwavering support during my internship. I am truly grateful for their constant encouragement, understanding, and motivation, which played a vital role in keeping me focused and inspired to give my utmost effort. Their presence has been invaluable, and I am incredibly fortunate to have them by my side.

Thank you,

Muhammed Hisham T

**TABLE OF CONTENTS**

1. **INTRODUCTION 6**
2. **DATA DESCRIPTION 6**
3. **PROBLEM STATEMENT 7**
4. **EXPLORATORY DATA ANALYSIS 7**
   1. **Sales Trends 7**
   2. **Product Performance 9**
   3. **Payment Mode Preference 10**
   4. **Lead Conversion Rates 11**
   5. **Coupon Code Usage 11**
   6. **Payment Processing Performance 12**
   7. **Market Channel Effectiveness 12**
5. **MODELING 14**
6. **DASHBOARD 16**
7. **RESULTS AND FURTHER WORK 17**

**INTRODUCTION**

The primary objective of this project was to employ a range of advanced Machine Learning techniques to analyze and make predictions based on real-time data. To achieve this, we obtained data from GUVI Geek Network Pvt. Ltd., a prominent company, which comprised the sales records spanning the past several months. By leveraging this dataset, we aimed to forecast future sales figures, providing valuable insights for informed decision-making.

Furthermore, this project sought to delve into the modern landscape of data visualization tools, particularly focusing on the utilization of PowerBI. Our aim was to explore how such tools are currently employed in the industry for data analysis purposes. PowerBI, renowned for its ease of use and learnability, proved to be a valuable asset throughout our analysis.

In addition to the predictive modeling and utilization of data visualization tools, we recognized the importance of thoroughly understanding the dataset. To this end, we conducted exploratory data analysis (EDA) to uncover underlying trends and patterns. By gaining a comprehensive understanding of the data, we were able to effectively enhance the accuracy and reliability of our sales forecasts.

In summary, this project encompassed multiple facets, including the application of Machine Learning techniques, the utilization of PowerBI for data visualization and analysis, as well as comprehensive exploratory data analysis. Through these efforts, we aimed to provide valuable insights into sales trends and future projections, ultimately aiding decision-makers in making informed and data-driven choices.

**DATA DESCRIPTION**

I obtained the sales data for a period of two months from a colleague who is currently employed at Guvi. The data is presented in CSV format and consists of eleven columns, namely: user ID, product code, payment status, coupon code, lead generation time, sales date, source, product amount including GST, payment mode, currency code, and transaction bank. In total, the dataset comprises 9514 rows and 11 columns.

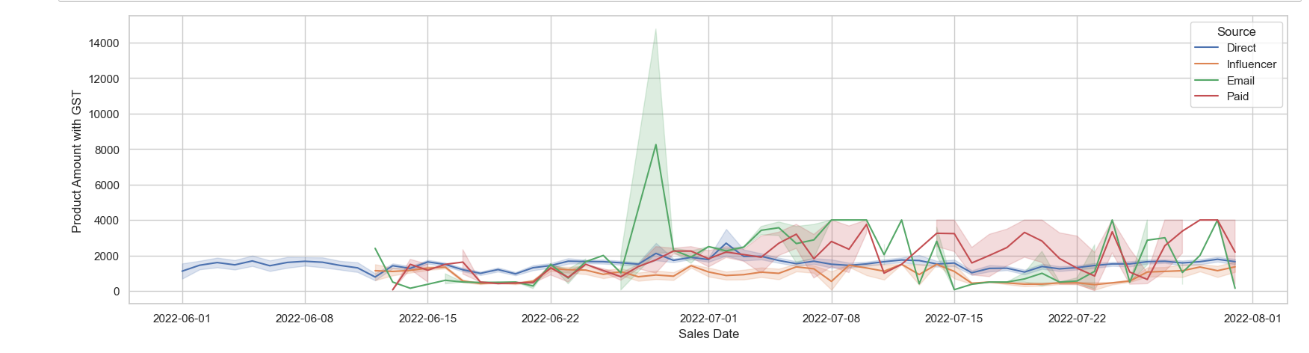
**PROBLEM STATEMENT**

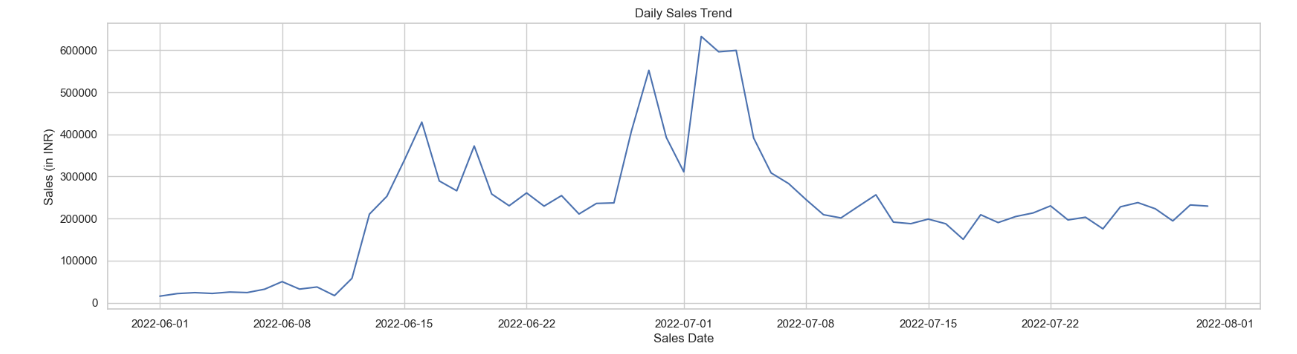
The primary objective is to forecast the sales of a company by utilizing data from past months. This entails analyzing historical sales data and employing statistical or machine-learning techniques to predict future sales trends. By examining patterns, seasonality, customer behavior, and other factors within the data, the aim is to generate accurate sales forecasts. Additionally, a Power BI dashboard will be developed to visualize sales data and provide insights based on various parameters. The dashboard will offer interactive visualizations, charts, and graphs that can be customized to showcase sales performance across different dimensions, such as time, product categories, regions, customer segments, and more. This will enable stakeholders to easily understand and analyze the sales data, identify trends, spot potential opportunities or issues, and make informed business decisions.

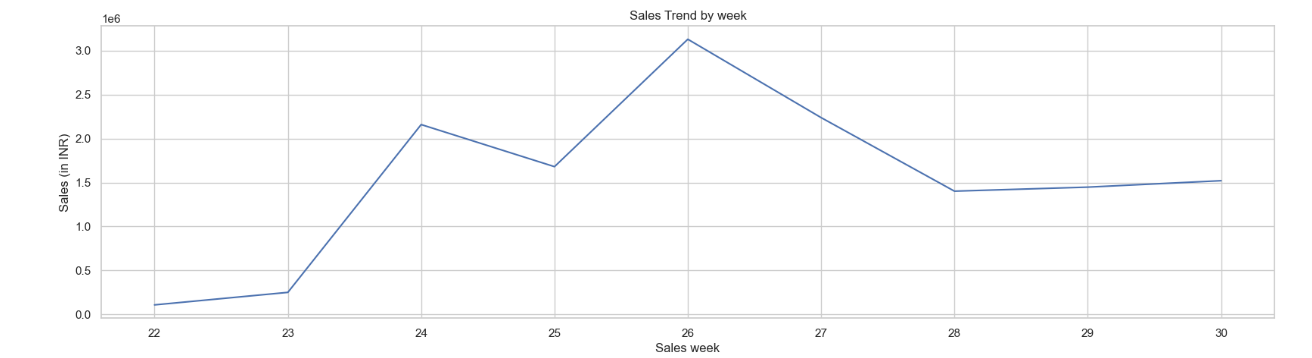
**EXPLORATORY DATA ANALYSIS**

**1. Sales trends**

By conducting an analysis of the 'Sales Date' column, one can discern the prevailing sales patterns that have emerged over the course of the previous two months. This comprehensive examination will enable a comprehensive understanding of the specific days of the week or particular periods within the month that exhibit the highest sales volumes.

****

****

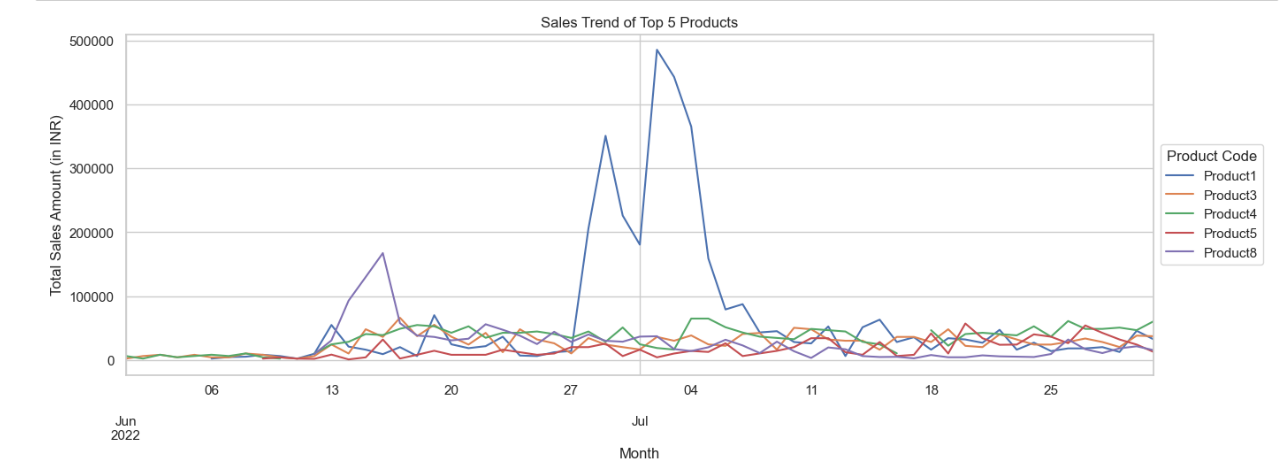
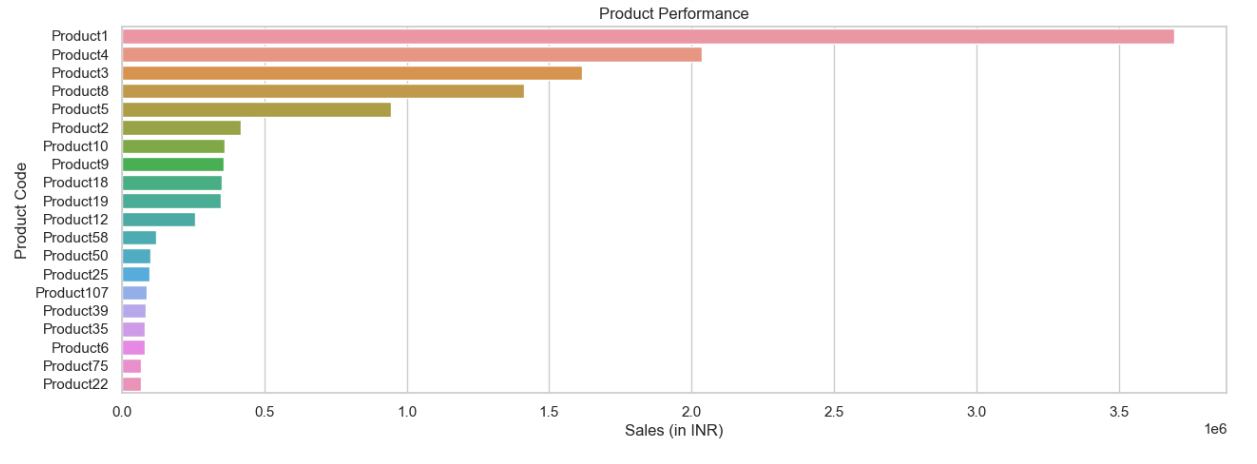
****

**Result:**

* During June 2022, the sales exhibited a progressive rise, specifically in the middle of the second week, indicating a gradual increase in sales volume over that period.
* Commencing from the middle of the first week of July 2022, the sales experienced a decline, signifying a decrease in sales figures during that particular timeframe.
* The introduction of additional sales channels, namely influencer marketing, email campaigns, and paid advertising, played a pivotal role in driving a notable increase in sales. This growth is particularly significant considering that prior to the utilization of these new sources, sales were solely dependent on a single channel, namely direct sales.
* Despite the decline in sales during the middle of the first week of July 2022, it is worth noting that the average sales margin for that period remained higher compared to the preceding month of June.

**2. Product performance**

Through an analysis of the 'Product Code' and 'Product Amount with GST' columns, we can effectively discern the products that are exhibiting strong performance and generating substantial revenue. This valuable insight will enable us to make well-informed decisions regarding prioritization of certain products while considering the possibility of discontinuing others**.**

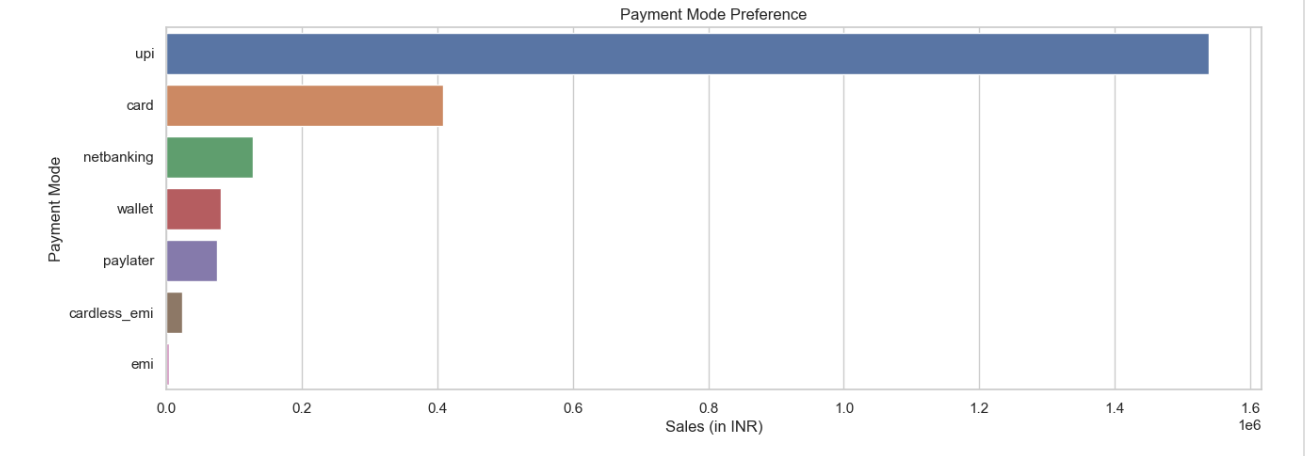
****

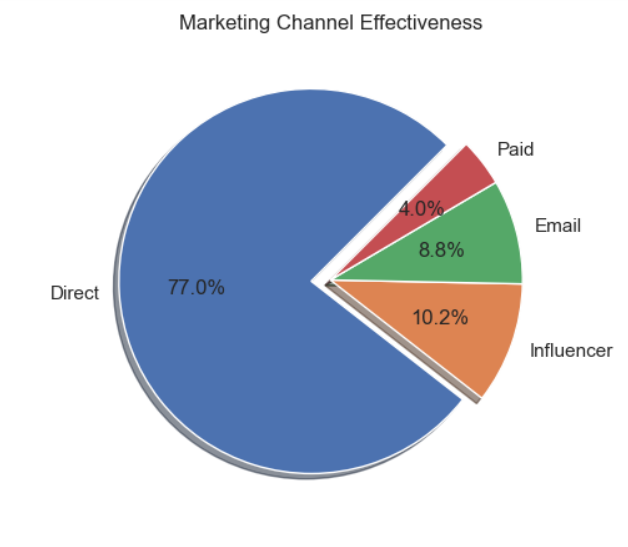
**Result:**

* The cumulative sales of Product1, Product4, Product3, and Product8 have surpassed the impressive milestone of generating a total revenue exceeding 10 lakhs.
* From the 27th of June to the 4th of July, Product1 has achieved a substantial level of market penetration and widespread popularity.

**3. Payment mode preference**

By thoroughly examining the data within the 'Payment Mode' column, we can gain valuable insights into the specific payment methods that customers favor, enabling us to strategically optimize our payment options in order to effectively cater to their unique needs and preferences.



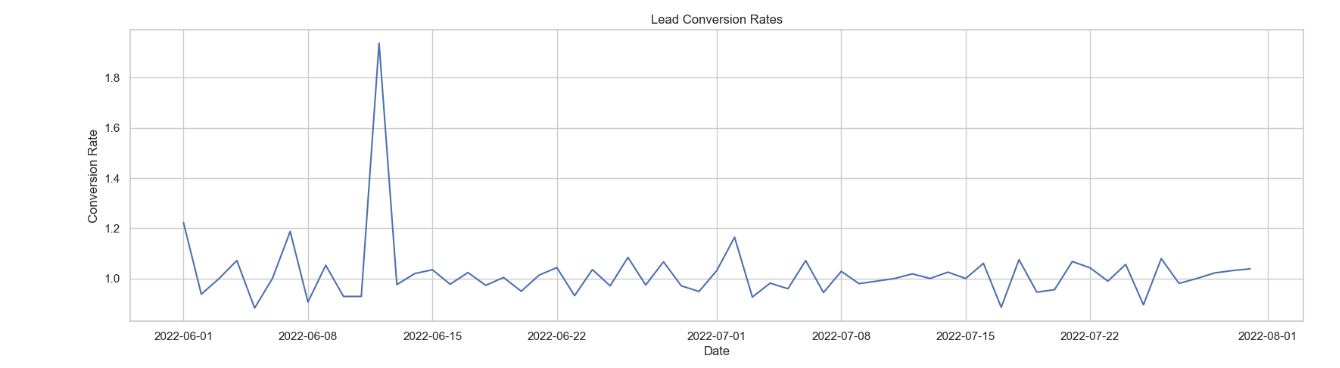


**Results:**

* Excluding the instances labeled as 7783 - 'Unknown Payment Mode', the majority of users have demonstrated a strong preference for conducting transactions through the widely favored payment methods of 'Upi' and 'card'.

**4. Lead conversion rates**

Through the process of comparing the data within the 'Lead Registered Time' and 'Sales Date' columns, it becomes possible to derive valuable insights regarding lead conversion rates, which provide a clear understanding of the average duration required for a lead to successfully convert into a sale. This analysis serves as a powerful tool for optimizing lead generation strategies and refining follow-up procedures, ultimately leading to enhanced conversion rates and overall business performance.

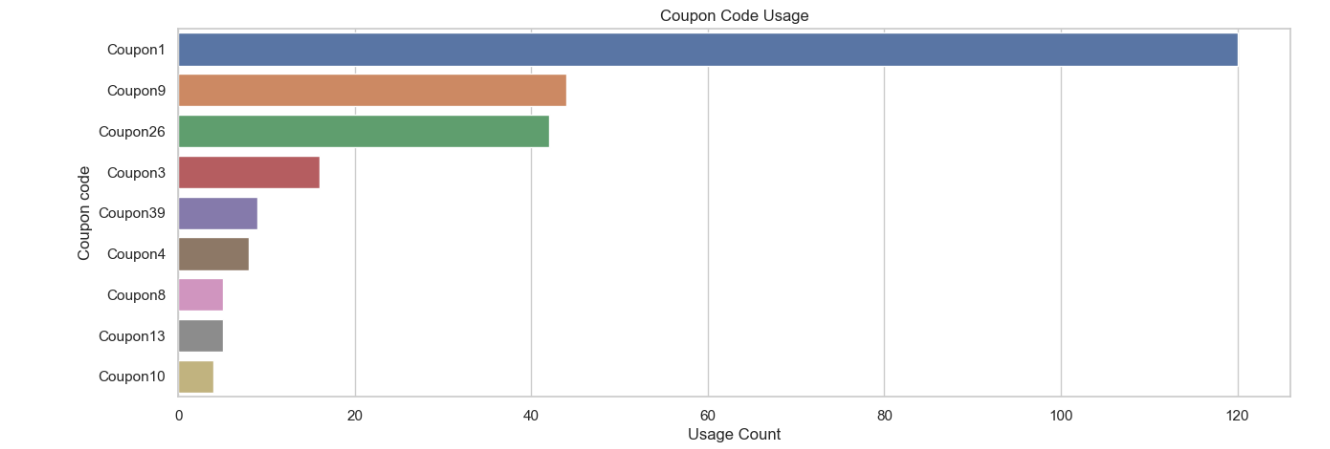


**Results:**

* During the middle of the second week in June 2022, the lead conversion time experienced a notable surge, which can be attributed to a sudden and significant upturn in sales.

**5. Coupon code usage**

Through comprehensive analysis of the data contained within the 'Coupon code' column, we can gain valuable insights into the frequency of customer coupon usage and determine which specific types of coupons are most widely favored. This analysis serves as a crucial tool in optimizing our coupon offerings, allowing us to effectively align with customer needs and preferences while effectively incentivizing sales.

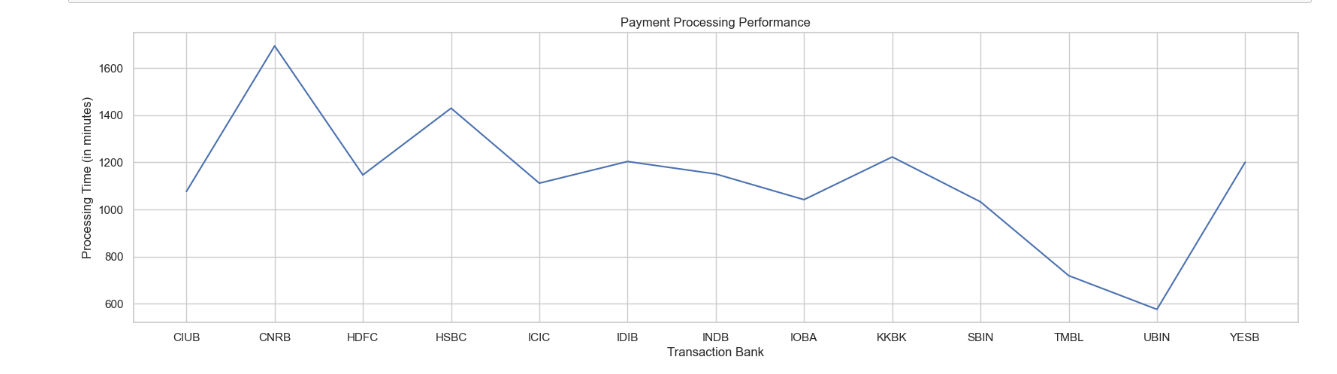


**Results:**

* In addition to encountering 8852 coupons with an 'unknown Coupon Status', there are higher usage counts for Coupon1, Coupon9, and Coupon26.

**6. Payment processing performance**

By leveraging the information provided in the 'Payment Status' and 'Transaction Bank' columns, businesses can effectively monitor the performance of their payment processing systems and promptly detect and address any sales-related issues or errors.

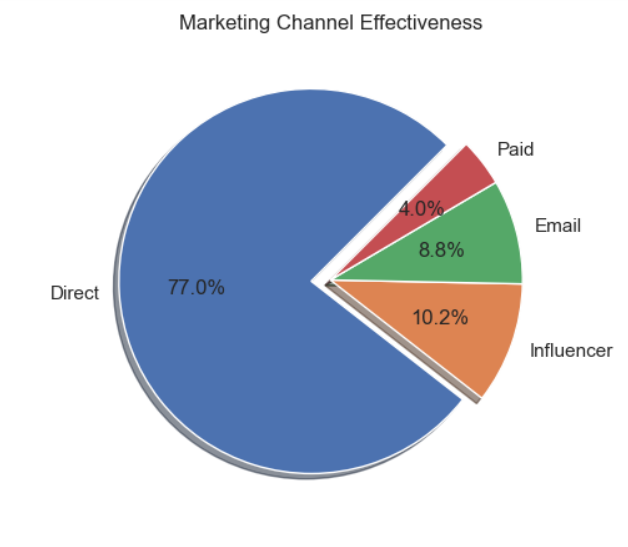
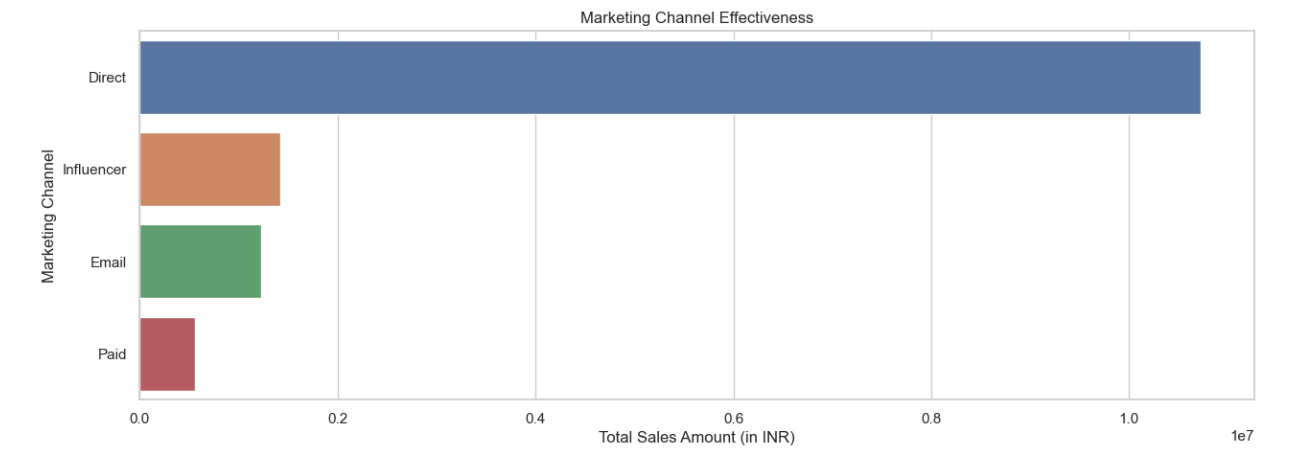


**Results:**

* Analyzing the graph reveals that the "CNRB" bank exhibits longer processing times compared to the "UBIN" bank, which demonstrates shorter processing times. It is worth noting that, excluding the 9421 instances categorized as 'Unknown Bank', these observations hold true.

**7. Marketing channel effectiveness**

The utilization of the 'Source' column enables businesses to gain insights into the primary marketing channels responsible for driving significant sales. This valuable information can be leveraged to optimize marketing strategies, ensuring that resources are directed towards the most effective channels, thereby maximizing the overall impact on the business.



**Results:**

* Over the course of the given Two Months, it can be observed that the marketing channel labeled "Direct" emerged as the most dominant driver of sales, exerting the greatest influence on the overall sales performance during that period.
* The influence of various sources, such as email campaigns and paid marketing efforts, has proven to be instrumental in driving notable growth in sales. These sources have played a substantial role in boosting sales figures, highlighting their significant impact on the overall success of the business.

**Note:** Overall sales in two months is 1.39 Crores.

Total Sales in each month: June - 57,88,559.64

July - 81,47,819.20

**MODELING**

**Linear Regression Model:**

Linear Regression is a supervised learning algorithm used to model the relationship between a dependent variable and one or more independent variables. In its simplest form, the mathematical expression for a linear regression model with a single independent variable is:

y = β₀ + β₁x + ε

Where:

y is the dependent variable (target variable) being predicted.

x is the independent variable (input feature) used to make predictions.

β₀ is the y-intercept (bias term) or the value of y when x is 0.

β₁ is the slope or coefficient that represents the change in y for a unit change in x.

ε is the error term representing the difference between the predicted and actual values of y.

The goal of linear regression is to estimate the values of β₀ and β₁ that minimize the sum of squared errors between the predicted values (ŷ) and the actual values (y). This is typically achieved using an optimization algorithm such as ordinary least squares (OLS).

For multiple linear regression, where there are multiple independent variables, the equation expands to: y = β₀ + β₁x₁ + β₂x₂ + ... + βₚxₚ + ε

Where:

x₁, x₂, ..., xₚ are the independent variables.

β₁, β₂, ..., βₚ are the coefficients corresponding to each independent variable.

The estimation of coefficients in linear regression involves solving a system of equations, and various numerical methods can be used, such as matrix operations or gradient descent optimization.

Once the coefficients are estimated, the model can be used to make predictions by substituting the input values into the equation.

Linear regression models can be extended to handle non-linear relationships by including higher-order terms or by applying transformations to the input variables. Regularization techniques such as Ridge Regression and Lasso Regression can also be applied to handle multicollinearity and prevent overfitting.

Overall, linear regression provides a straightforward and interpretable approach to modeling relationships between variables and making predictions in machine learning.

**IMPLEMENTATION OF THE MODEL**

Step 1: Converted the "Sales Date" column to Datetime.

Step 2: Calculate total sales in each month.

Step 3: Train a regression model. Split into test and train data.

Step 4: Prepared next month's data in the form of a DataFrame and predicted the sales for it.

Step 5: Similarly, prepared the data for more than two months in the form of the DataFrame and predicted the sales for the same.

In this way, we got the predicted sales for the next three months.

The sales for the next three months are given below:

2022 – 08(August): 10507078

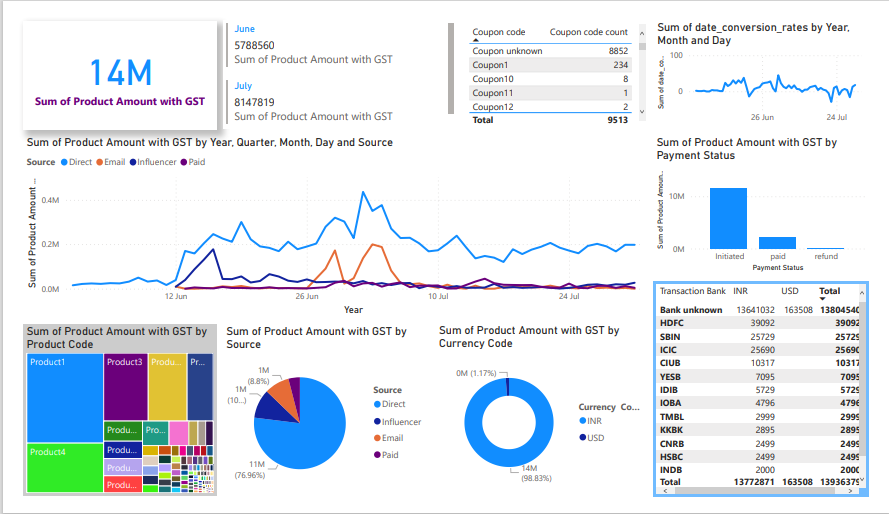
2022 – 09(September): 12866338

2022 – 10(November): 15225597

**DASHBOARD**

For dashboard creation, I used the PowerBI tool. Power BI is a business analytics tool developed by Microsoft that enables users to visualize and analyze data from various sources. One of the key features of Power BI is the ability to create interactive and dynamic dashboards. Dashboards provide a consolidated view of important data and insights, allowing users to monitor key metrics, track performance, and make data-driven decisions.

Power BI dashboards offer a powerful way to visualize, analyze, and monitor data in an interactive and dynamic manner. They enable users to gain insights, track performance, and make data-driven decisions. With its user-friendly interface and robust features, Power BI has become a popular choice for organizations seeking to leverage their data effectively.



The comprehensive dashboard incorporated and visualized a wealth of transaction details, offering users the flexibility to choose from a variety of options such as selecting specific months, coupon codes, products, transaction banks, and more, thereby enhancing the customization and analytical capabilities for thorough data exploration and analysis.

**RESULTS AND FURTHER SCOPE**

The primary objective of the project revolved around utilizing machine learning techniques, specifically prediction or forecasting, to estimate company sales by leveraging historical data from previous months. This initiative aimed to showcase the effectiveness of applying these predictive models to real-time data, highlighting their potential in generating valuable insights. Additionally, the project incorporated the creation of a Power BI dashboard to demonstrate how data analysis can be efficiently conducted using advanced visualization tools, exemplifying the power and utility of tools like Power BI in facilitating comprehensive data analysis.

To further enhance the analysis, the data can be subjected to an expanded range of supervised and unsupervised machine learning algorithms, enabling the extraction of additional insights and patterns. Additionally, employing a wider array of statistical tools for exploratory data analysis (EDA) could yield even more comprehensive findings. The dataset holds the potential for unveiling numerous trends and valuable information that can contribute to a deeper understanding of the underlying patterns and dynamics present within the data.