|  |
| --- |
|  |
| Data Science Project Report  Sales Data Analysis and Forecasting/Prediction |
| |  |  |  | | --- | --- | --- | | **Mukand Krishna 20K-0409** | **12/31/23** | **Data Science** | |

Section 1: Introduction:

This Sales dataset comprises sales information, including product details, quantities, and financial aspects. The goal is to provide a comprehensive analysis, uncovering patterns and trends that can inform strategic business decisions.

Data Overview:

The dataset encompasses **290,514 entries and 12 columns**.

Key columns include:

* DistributorCode & DistributorName: Unique identifier and name of the distributor.
* ClientCode & ClientName: Unique identifier and name of the client or store.
* BrickName: A detailed location or category identifier.
* Product & SKU: Product details, with SKU offering a granular identifier.
* InvoiceDate: Date of the transaction.
* Units, Bonus, Discount: Quantities and financial aspects of the transaction.
* ValueNp (Net Profit): Total revenue from sales.



Section 2: Exploratory Data Analysis

1. **Product Wise Analysis:**

A graph of different colored bars

Description automatically generated

* Products Distribution in Cities:

A graph of different colored bars

Description automatically generated

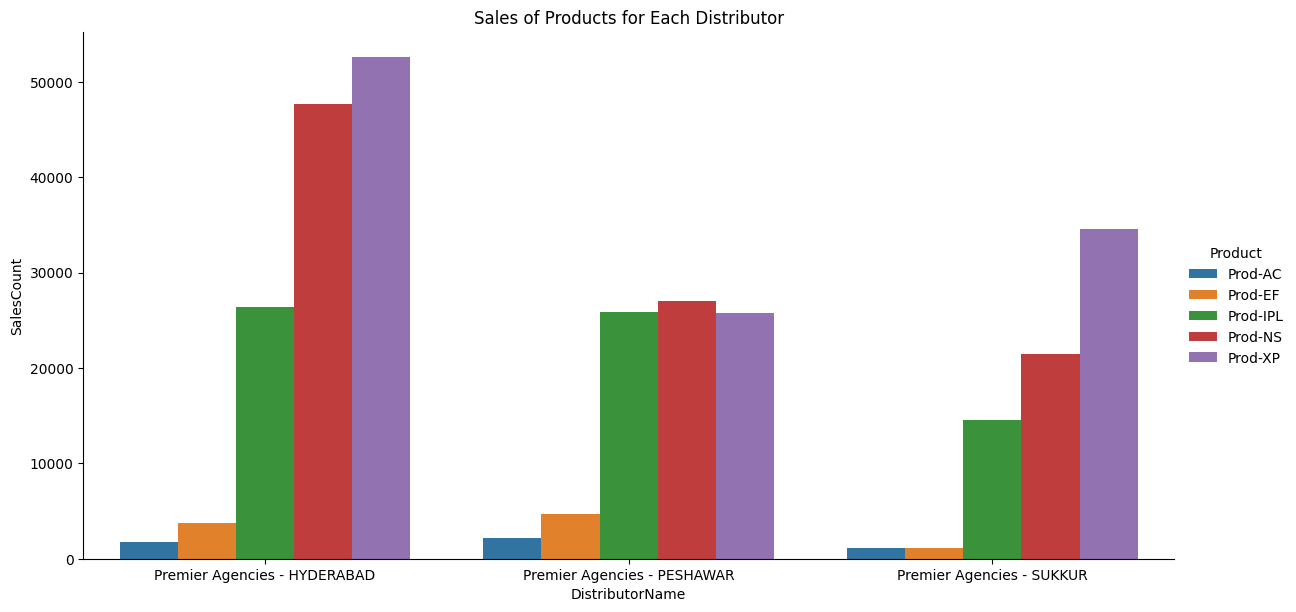
* Products Distribution in Bricks:

A screenshot of a computer

Description automatically generated

1. **Distributor Wise Analysis:**

* Product Sales for Each Distributor:



* Brick Wise Analysis for Each Distributor:

A screenshot of a data analysis

Description automatically generated

* Yearly Sales for Each Distributor:

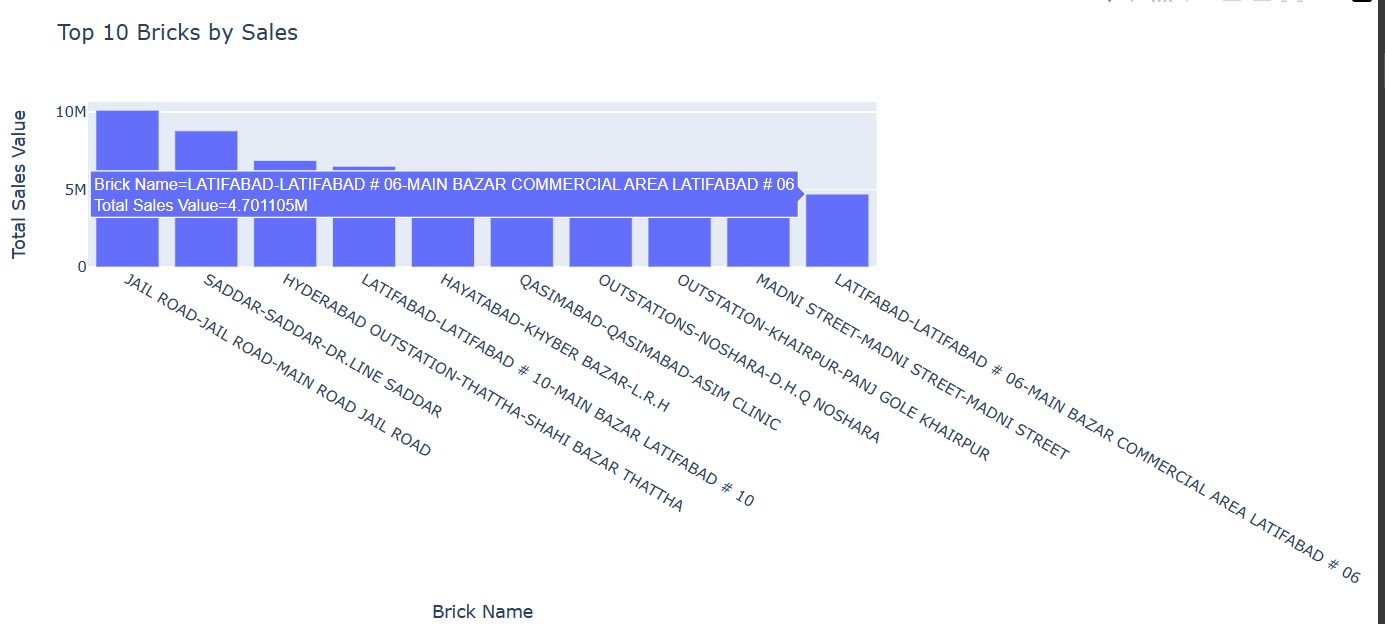
A graph of a bar chart

Description automatically generated with medium confidence

*'Premier Agencies HYDERABAD' had a substantial increase in sales from 2016 to 2017.*

1. **Brick Analysis:**

* Top 10 Bricks by Sales:



1. **Yearly Sales Trends for Each City:**

A graph with lines and numbers

Description automatically generated with medium confidence

Section 3: Data Preprocessing:

1. **Handling DateTime:**

Began by converting the 'InvoiceDate' column to a datetime format to facilitate temporal analysis.

1. **Time Series Analysis:**

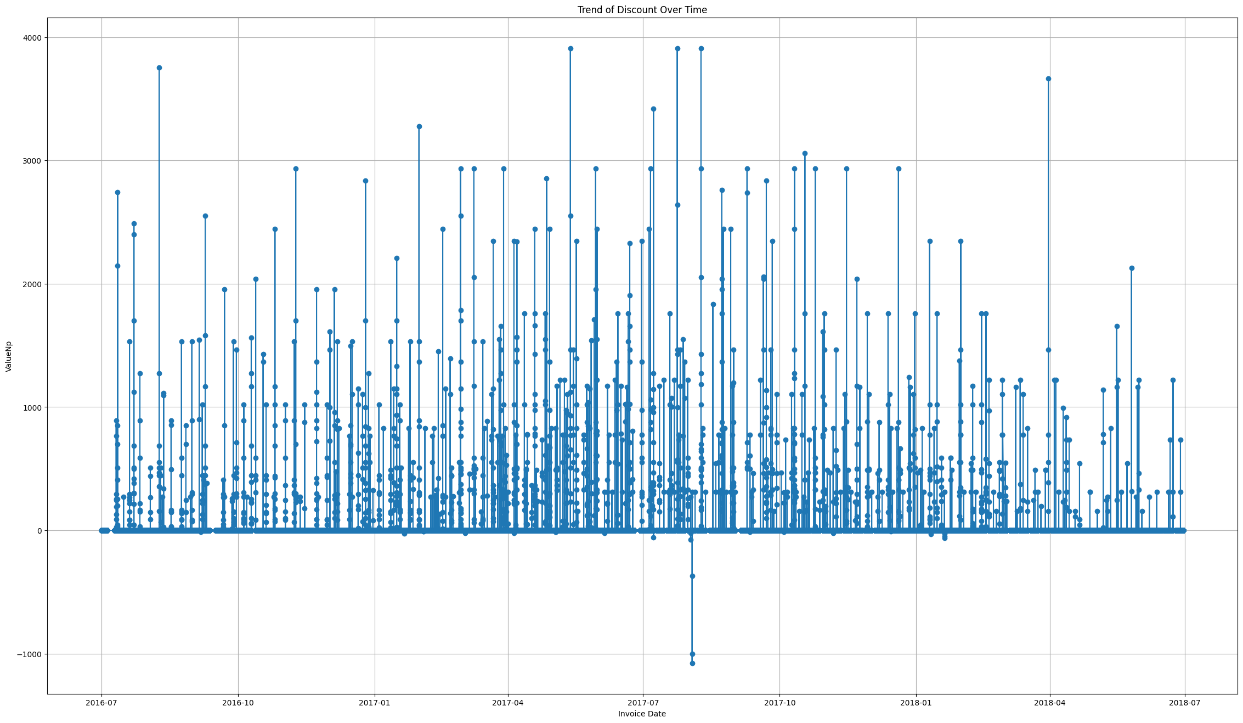
Plotted the trends of 'ValueNp,' 'Discount,' and 'Units' over time to identify any patterns or anomalies.

* Trend of ValueNp Over Time reveals overall sales trends.

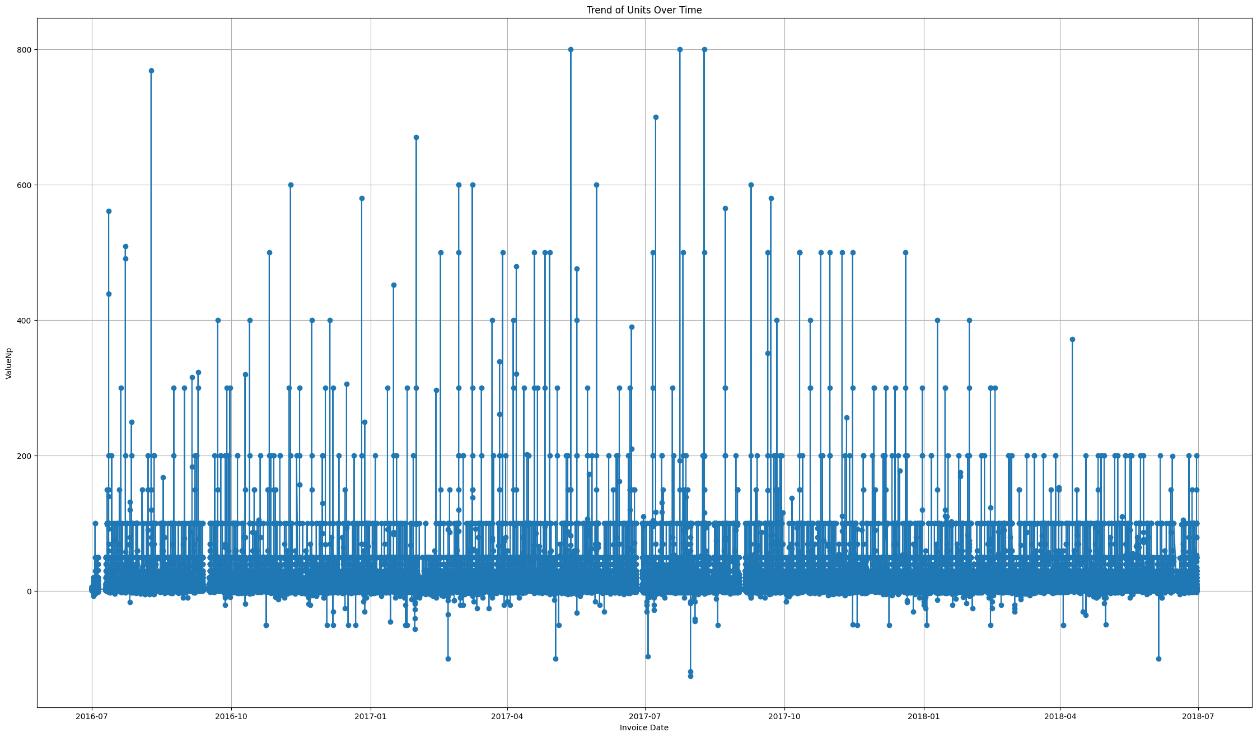
A blue lines on a white background

Description automatically generated

* Trend of Discount Over Time showcases variations in discount patterns.



* Trend of Units Over Time illustrates fluctuations in product units sold.



1. **Seasonal Decomposition**:

We performed seasonal decomposition using the additive model to identify underlying trends, seasonality, and residuals. The decomposition helps discern patterns in the data.

A screenshot of a graph

Description automatically generated

1. **Monthly Sales Analysis:**

* Monthly sales were analyzed to uncover patterns and trends over the two-year period.

A graph with lines and dots

Description automatically generated

A colorful circle with numbers and numbers

Description automatically generated

Section 4: Feature Engineering for Prediction Purpose:

1. **Min-Max Scaling:**

Applied Min-Max scaling to three columns **('Units,' 'Discount,' 'ValueNp')** to bring them to a standard scale. Visualizations of the original and scaled distributions are presented.

A screenshot of a graph

Description automatically generated

A group of black lines

Description automatically generated with medium confidence

1. **Encoding Categorical Features:**

Categorical features such as 'DistributorName,' 'Product,' 'SKU,' 'ClientName,' and 'BrickName' were encoded for model compatibility.

A screenshot of a computer program

Description automatically generated

1. **Drop Unnecessary Columns and Final Dataset:**

Columns like 'InvoiceDate,' 'ValueNp,' 'Discount,' and 'Units' were dropped to streamline the dataset for modeling purposes.

A screenshot of a computer program

Description automatically generated

The dataset is now ready for **model training and evaluation**.

Section 5: Predictive Modeling:

The predictive modeling phase aimed to forecast future sales using a robust ensemble model. Utilized a **Voting Regressor combining RandomForest and GradientBoosting models**.

The model was trained on *80% dataset*, *tested on 20% dataset*, and evaluated using metrics such as *Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared.* The evaluation results indicated high accuracy and reliability.

Model Evaluation Results:

A screenshot of a computer error

Description automatically generated

The predictive model's accuracy was further visualized through a scatter plot comparing actual vs. predicted values.

A graph with red and blue dots

Description automatically generated

**Predicted and Actual Scaled ValueNp values ( 20%):**

A screenshot of a computer screen

Description automatically generated

Section 6: SHAP Interpretation:

Employed SHAP values to interpret the model's decisions and understand the importance of features in predicting sales.

SHAP summary plots provided a global interpretation, while dependence plots showcased the impact of individual features on predictions. They helped in identifying key factors influencing sales predictions and insights into the model's decision-making process.

* **SHAP Summary Plot:**

A graph with numbers and lines

Description automatically generated with medium confidence

Plot describes that **Units** has major impact followed by Product Types i.e., SKU

* **SHAP Decision Plot:**

A graph with numbers and lines

Description automatically generated

**SKU (Stock Keeping Unit)** **Product**

A graph of red and blue dots

Description automatically generatedA graph of red and blue lines

Description automatically generated

A graph of a number of dots

Description automatically generated with medium confidence

**Brick Name**

**Client Name(Encoded) Distributor Code**

A graph with red and blue lines

Description automatically generatedA graph with red and blue dots

Description automatically generated

Section 7: Conclusion:

* What We Explored:

We dug into a big set of sales data to understand how things work. It's like looking at sales trends, figuring out when sales are high or low, and finding patterns over time.

* What We Found:

We noticed interesting things in the data, like when sales go up or down. We also checked out what happens each month and saw which months bring in the most sales.

* What We Predicted:

We used cool math stuff to predict future sales. Our predictions were quite accurate, and we measured this accuracy using numbers like MSE, MAE, and R-squared.

* How We Figured It Out:

We made a smart model that learned from the data. It's like having a super-smart friend who can predict what might happen next based on what happened before.

* Understanding the Predictions:

We didn't stop at predictions; we also wanted to know why the model made those predictions. We uncovered the secrets behind the predictions by looking at important factors like the distributor, product, and client names.

* Why It Matters:

All this isn't just about numbers; it's about helping businesses make smart decisions. By understanding sales patterns and predicting the future, we're handing over a powerful tool for making wise choices.

* The Big Picture:

In wrapping up, our journey was like solving a puzzle. We looked at data, predicted the future, and understood why. It's not just about numbers; it's about empowering businesses to make informed decisions.

**End of Project — Unraveling Sales Secrets for Smart Decisions!**