This research aims to analyze retail sales data to understand the factors that contribute to sales performance and to develop a machine learning model capable of predicting sales values accurately. The study focuses on identifying the relationships between product categories, regional distribution, logistical factors, and temporal trends, and how these variables collectively influence sales outcomes. By combining exploratory data analysis and predictive modeling, this research seeks to bridge analytical insights with practical business applications.

The initial phase of the study involves exploratory data analysis to examine sales distribution patterns across different product categories and regions. The analysis reveals that sales performance varies significantly among product categories, indicating differences in customer demand and purchasing behavior. Regional analysis further shows uneven sales distribution across geographical areas, highlighting the importance of location-based market strategies. Additionally, temporal analysis of monthly sales trends uncovers seasonal patterns, suggesting that certain periods experience higher demand than others. Logistical factors, particularly shipping duration, are also identified as influential variables that may impact customer satisfaction and sales volume.

To enhance the analytical findings, a machine learning regression approach is employed to predict sales values based on the available features. A Gradient Boosting Regressor is selected due to its ability to model complex and non-linear relationships within the data. Feature engineering techniques are applied to transform categorical variables and extract meaningful numerical representations from date-related attributes. The dataset is then split into training and testing subsets to ensure proper model evaluation.

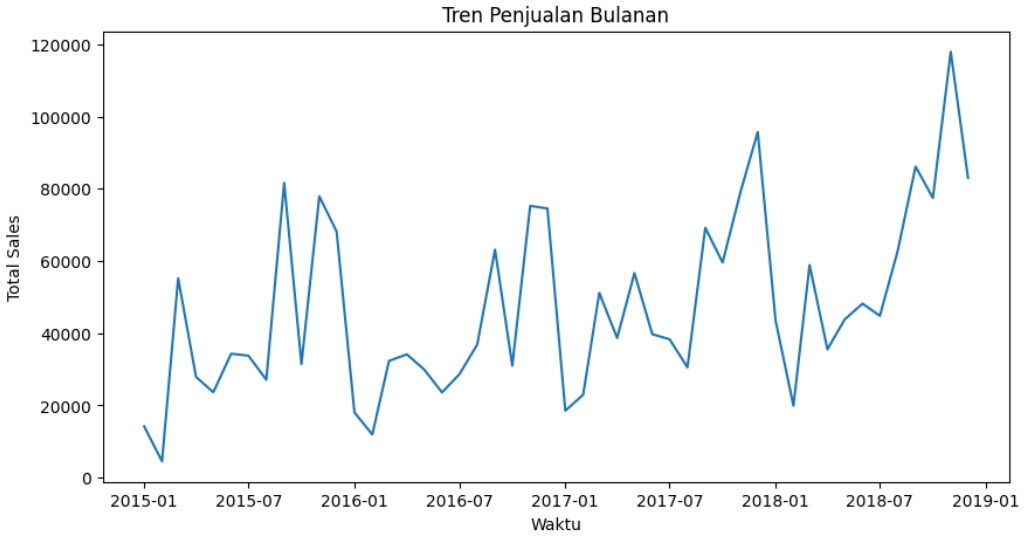
MAE	183.26
RMSE	333.54
R^2	0.7154

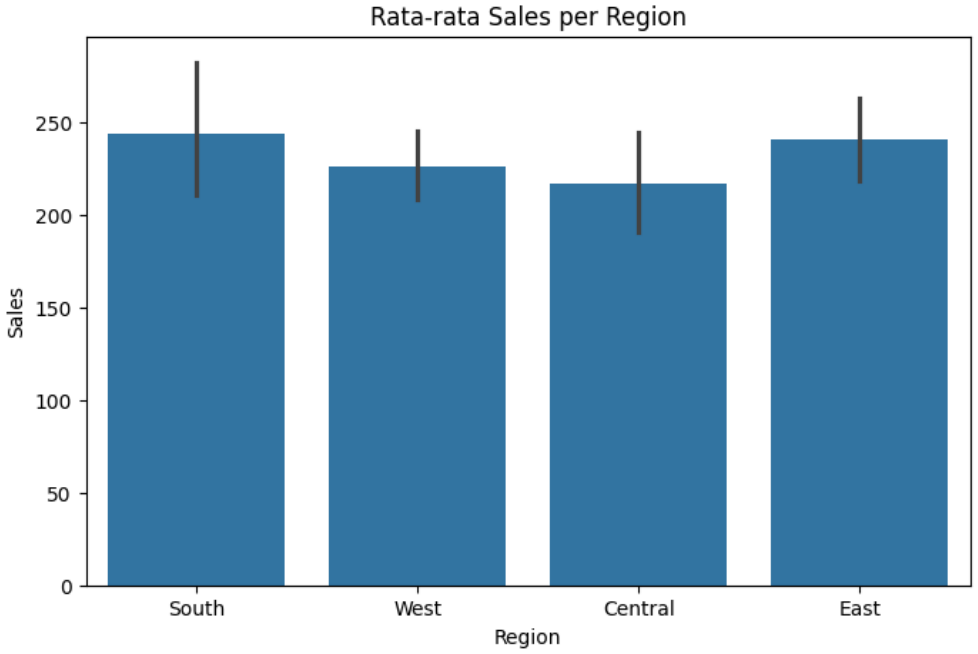
The predictive model achieves an R^2 score of 0.71, indicating a strong level of explanatory power and predictive performance. This result demonstrates that the model is able to capture key patterns in historical sales data and generate reliable predictions. The evaluation confirms that variables such as product category, region, shipping duration, and temporal features play a significant role in determining sales outcomes.

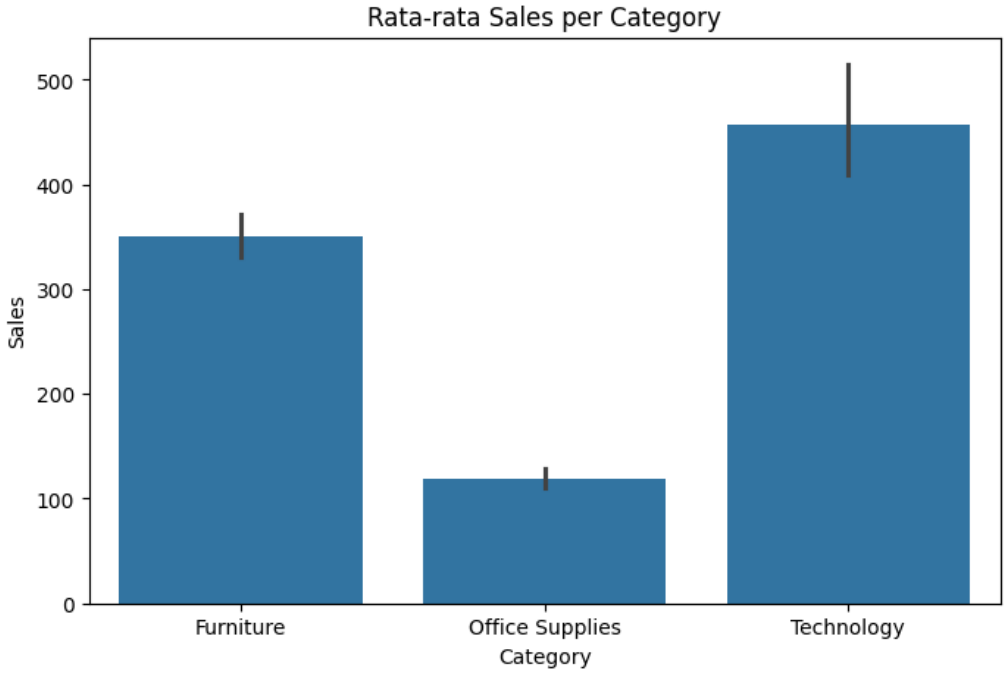
From a business perspective, the findings of this research provide valuable insights that can support strategic decision-making. The analysis enables businesses to focus marketing and promotional efforts on high-performing product categories and regions, optimize logistics and shipping processes to improve sales performance, and leverage seasonal trends for better demand forecasting and inventory planning. The developed machine learning model can serve as a decision-support tool for sales forecasting, operational optimization, and long-term business planning.

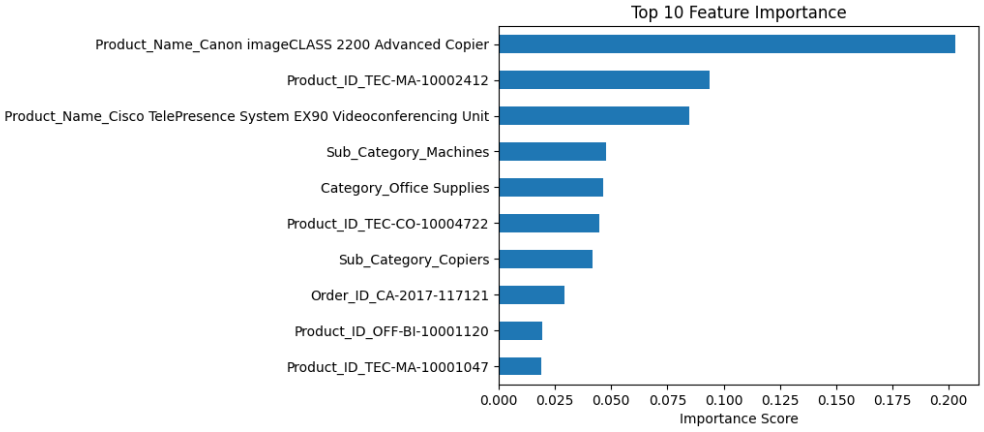
In conclusion, this research demonstrates how data-driven analysis and machine learning techniques can be effectively applied to retail sales data to generate actionable insights and predictive capabilities. The integration of exploratory analysis and predictive modeling highlights the role of data science in supporting both analytical understanding and practical business strategies.

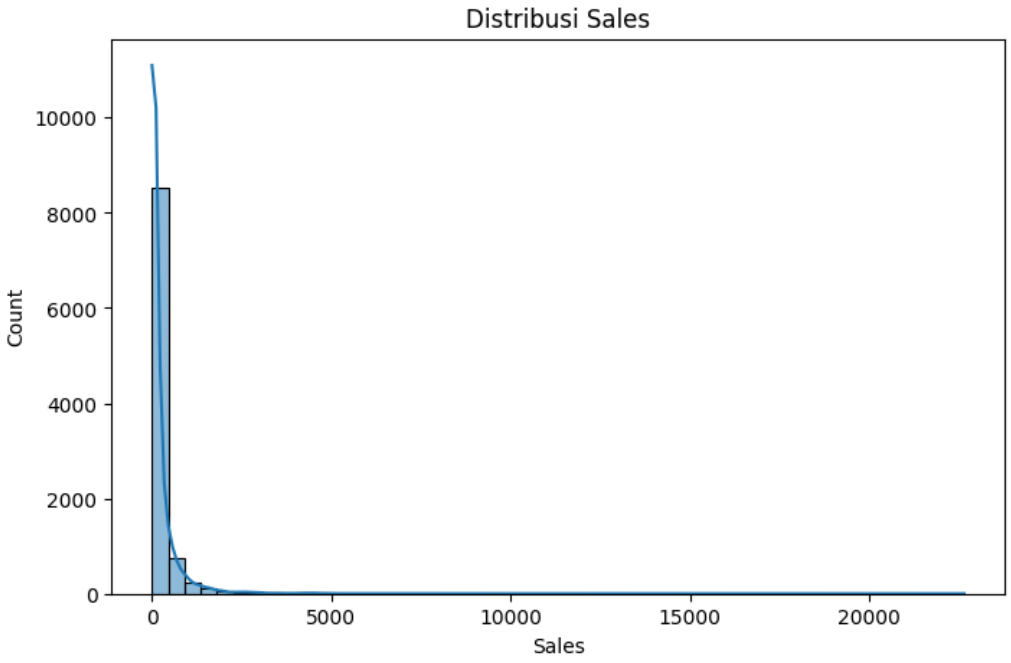
Attachment :


Tren Penjualan Bulanan
Total Sales
Waktu
0
20000
40000
60000
80000
100000
120000
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
2018-07
2019-01


Rata-rata Sales per Region
Sales
Region
0
50
100
150
200
250
South
West
Central
East


Rata-rata Sales per Category
Sales
0
100
200
300
400
500
Furniture
Office Supplies
Technology
Category


Top 10 Feature Importance
Product_Name_Canon imageCLASS 2200 Advanced Copier
Product_ID_TEC-MA-10002412
Product_Name_Cisco TelePresence System EX90 Videoconferencing Unit
Sub_Category_Machines
Category_Office Supplies
Product_ID_TEC-CO-10004722
Sub_Category_Copiers
Order_ID_CA-2017-117121
Product_ID_OFF-BI-10001120
Product_ID_TEC-MA-10001047
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Importance Score


Distribusi Sales
Count
10000
8000
6000
4000
2000
0
0
5000
10000
15000
20000
Sales