

Chavaporn T.

Links: https://github.com/Pekky28/Portfolio1-PredictSale.gi

OutLine

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix



Executive Summary Summary of Methodology

The project commences with data collection from two primary sources: insights provided by the owner and sales data from three nail salon branches—Prompong, Sukhumvit, and Thonglor.

After data collection, a data cleaning process is applied to ensure data consistency and accuracy. This includes handling missing values, correcting misspelled column headers, and removing irrelevant rows.

Exploratory Data Analysis (EDA) is conducted to gain a deep understanding of the collected data. Insights are derived through statistical summaries, visualizations, and exploratory techniques.

The next step involves building a predictive model aimed at forecasting next month's total sales for each branch and the overall total sales across all branches. Machine learning algorithms and statistical techniques will be applied to the engineered features, and model performance will be evaluated using appropriate methods. Once the model is validated, it can be deployed to estimate the likelihood of total sales and predict whether the outcome will be positive or negative. This estimation can be used for marketing strategies, sales forecasting for the next month (based on limited historical data), and data-driven decision-making for each nail salon branch.

In summary, the methodology combines data collected from a three-month historical dataset for each branch and insights from the owner. It encompasses data cleaning, exploratory data analysis, feature engineering, and predictive modeling to achieve the project's objective of predicting the likelihood of total sales for each branch and the total sales across all branches. This comprehensive approach ensures that the model is well-informed by critical data sources and facilitates data-driven decision-making.

Executive Summary Summary of Result

Throughout the project, a comprehensive analysis of data from three nail salon branches was conducted. This analysis encompassed various aspects of sales, employee performance, and predictive models. The insights and models derived from this analysis can be valuable for nail salons seeking to optimize future sales and develop effective marketing strategies for improvement. Overall, the project demonstrated how data analysis and predictive modeling can provide valuable insights into retail businesses like nail salons, potentially influencing strategic decisions and enhancing sales.



Introduction

Project Background Context

This project primarily focuses on predicting the total sales of each nail salon branch, a critical factor in running the business and implementing marketing strategies. The project encompasses various phases, including data collection, data cleaning, exploratory data analysis (EDA), and predictive modeling.

The goal is to gain insights into factors such as how each employee performs in terms of sales for each branch, predict the likelihood of outcomes, estimate sales, and explore other variables that can influence total sales. Additionally, the project explores how predictive models can be used to reasonably estimate sales.

The results aim to assist in making informed decisions regarding resource allocation, employee performance improvement, and strategic adjustments to enhance overall sales and business success.

Problems/Objective

How can we accurately predict the total sales for each of the three nail salon branches?

Can we reliably estimate the likelihood of various sales outcomes for the nail salon branches?

How can predictive models be effectively leveraged to reasonably estimate future sales and assist in data-driven decision-making for the nail salon branches?

Methodology

Data Collection Methodology

Perform data wrangling

Perform EDA using visualization

Perform predictive analysis using regression models





Data Collection

Start: The data collection process begins.

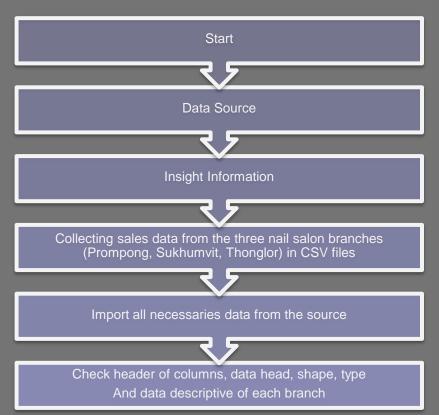
Data Sources: Data is sourced from the owner of the nail salon shops, including CSV files containing historical datasets for each of the three branches:

This data encompasses information on sales, employee performance (including employee-customer interactions per day), and daily total sales.

Data Import: All necessary data is imported from the source files. Subsequently, the headers of columns, initial data samples, data shape, and data types are checked.

Data Description: Descriptive statistics and characteristics of the data for each branch are examined.

Key Phases: Important data types within this process include "Total Sales," "Employee per Customer," and "Historical Records."



Data Wrangling

The data wrangling process begins by importing the necessary libraries for data manipulation and analysis.

Next, the datasets for the three nail shop branches are loaded, and initial exploration is conducted to understand their structure and content.

We check the names and values of the variables to get an overall picture of the data.

The analysis includes examining the column headers, data types, and identifying missing values that require handling for further analysis and modeling.

If necessary, we may convert the 'Date' column to a numerical format and categorize the dataset for each branch, considering the data for the past three months.

Additionally, we might combine all datasets for the purpose of analyzing overall predictions.

The ultimate goal of data wrangling is to prepare the data for downstream tasks.

including data analysis, machine learning, and visualization.

Start



Importing Libraries:

pandas: For data manipulation and analysis. numpy: For numerical operations. matplotlib or seaborn: For data visualization



Loading Data: Use pd.read_csv() to load the CSV files for each branch Prompong, Sukhumvit, and Thonglor



Data Cleaning:

Handle missing values using 'pd.fillna()' or 'pd.dropna()

Correct misspelled column names with 'pd.rename()'

Correct misspelled column names with 'pd.drop()'

EDA with Data Visualization

Bar Chart: Used for visualizing the total sales and total customers of employees all across 3 branches. Bar charts are effective for comparing and displaying categorical data, such as performance of employees.

Line Chart: Used to visualize the prediction the total sale of each branch Prompong, Sukhumvit, and Thonglor for the next month. Line charts are useful for displaying trends and changes in data over time or across a continuous variable

Scatter Plot: scatter plots are effective tools for visualizing and analyzing the relationship between two continuous variables, such as the day of the month and total sales, and are particularly useful for identifying patterns, trends, and outliers in the data. It is important for detecting unusual sales figures on certain days.



Summary of Predictive Analysis Using ARIMA Model and Random Forest Regression

Why These Models Were Selected for the Project:

ARIMA Model Selection:

Rationale: ARIMA is chosen when historical sales data show time-dependent patterns such as seasonal variations or autocorrelation.

This is relevant for businesses with recurring sales patterns.

Project Application: In the project, the ARIMA model is applied to understand and forecast short-term sales fluctuations based on historical data, providing insights into trends and seasonality.

Random Forest Regression Selection:

Rationale: Random Forest Regression is selected when total sales are influenced by multiple factors that have complex, non-linear relationships. This model can handle both numerical and categorical variables, which is common in sales prediction scenarios.

Project Application: In the project, Random Forest Regression is used to consider the influence of various factors, such as employee performance on total sales. It helps capture and provides insights into how different variables impact sales.



Summary of Predictive Analysis Using ARIMA Model and Random Forest Regression

The choice of these models is driven by the nature of the data and the complexity of relationships within project.

ARIMA focuses on time-dependent patterns, while Random Forest Regression offers versatility in capturing complex and varied influencing factors on total sales. Together, these models provide a holistic approach to sales prediction, accounting for different types of influencing factors and historical sales trend sales.



Insight Draw From EDA

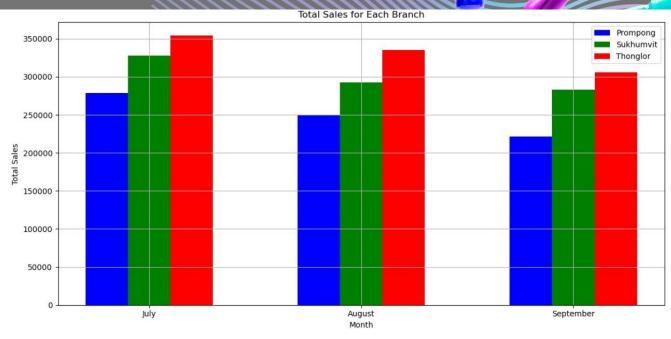


Total Sales For Each Branch 3 months

(July-Sep 2023)

Total Sales for Each Branch in the Last 3 Months (July to September 2023):

The overall performance indicates that the Prompong branch has the lowest total sales across all three branches, while Thonglor has the highest total sales, followed by Sukhumvit. Additionally, the total sales across all three branches have shown a decline from July to the end of September.

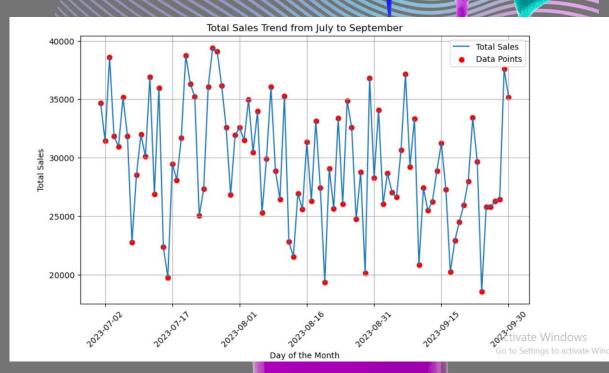


Total Sales Trend Of All Branches (July-Sep 2023)

The line graph illustrates the total sales trends of all three branches over time. The graph shows that all three branches performed well from around July 2nd to approximately July 10th before experiencing a rapid decline around July 15th. After July 17th, the graph saw a significant increase, reaching its highest peak during July 17th to August 1st, with total sales exceeding 35.000 baht.

At the beginning of August, the total sales maintained a relatively consistent volume, ranging from approximately over 25,000 to 35,000 baht. However, there was a sudden drop just before August 16th, and the total sales continued to fluctuate, with values ranging from 20,000 to just under 40,000 baht until August 31st.

Beginning of September, the graph remained unstable, with values fluctuating between over 25,000 and approximately 37,000 baht. After September 15th, there was a significant drop in total sales, which gradually began to rise around September 20th but experienced another drop around September 25th. However, after September 25th, there was a significant increase in total sales, which continued until the end of September.



Relationship between day and total sale of the day across 3 months from all branches(July-Sep 2023)

Sales per branch per day:

Days 1-5: Sales per branch range from 6,000 to 10,000 Baht. Some branches have the opportunity to exceed 15,000, especially around day 3.

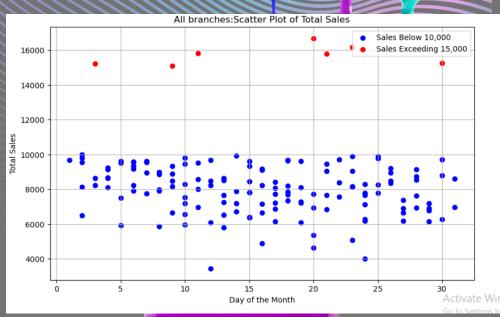
Days 6-10: Sales per branch range from 6,000 to 10,000 Baht. Some branches have the opportunity to exceed 15,000, especially around day 9.

Days 11-15: Sales per branch range from 6,000 to 10,000 Baht. Some branches experience a significant drop, going below 4,000 Baht. Some branches have the opportunity to exceed 15,000, especially around day 11.

Days 16-19: Sales per branch range from 5,000 to 9,000 Baht. No branch exceeds 15,000 in sales during this period.

Days 20-25: Sales per branch range from 4,000 to 10,000 Baht. Some branches have the opportunity to exceed 15,000.

Days 26-30: Sales per branch range from 6,000 to 10,000 Baht. Some branches have the opportunity to exceed 15,000, particularly on day 30.



Prompong Branch July-SEP 2023 📶

From July 2nd to July 17th, the sales volume ranged from approximately over 6,000 to 14,000 baht. However, there was a drastic drop in sales around July 20th, with values falling under 6,000 baht.

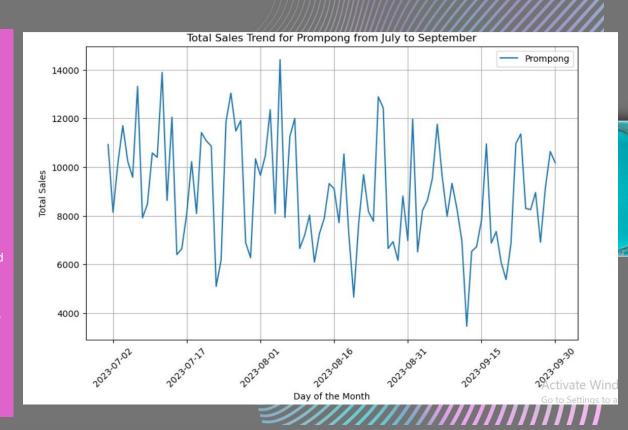
The sales then increased around July 25th, exceeding 12,000 baht, but experienced another significant drop just before the end of July.

The sales increased again just before August 1st. During the first half of August, from August 1st to August 16th, the sales ranged between 6,000 and over 14,000 baht.

However, there was a significant drop in sales after August 16th, with values falling below 6,000 baht. After this drop, the sales rapidly increased and maintained values between 6,000 and approximately 13,000 baht for the rest of the month, from August 16th to August 31st.

In September, the graph continued to fluctuate, with sales ranging from 4,000 to just under 12,000 baht, but not exceeding 12,000 baht.

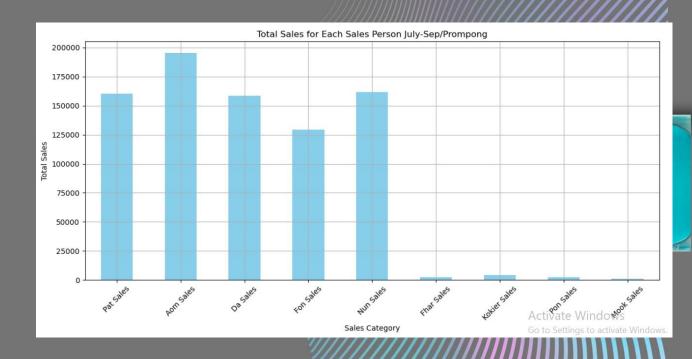
After September 15th, the graph showed a rapid rise, and the sales volume remained between approximately over 5,000 and not exceeding 12,000 baht until the end of September



Prompong Branch July-Sep

All employees' performance on total sales ranges between over 125,000 and 175,000 baht. The highest total sales are achieved by Aom, with the performance of other employees being relatively close.

Aom's sales are more prominent than those of other employees, followed by Nun and Pat, respectively.

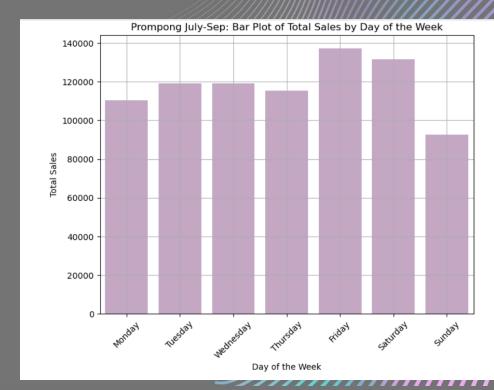




Prompong Branch July-Sep

Prompong: Bar Plot of Total Sales by Day of the Week

Display a bar plot comparing the days of the week across three months. Friday has the highest chance of achieving the highest sales, while Sunday has the lowest chance.



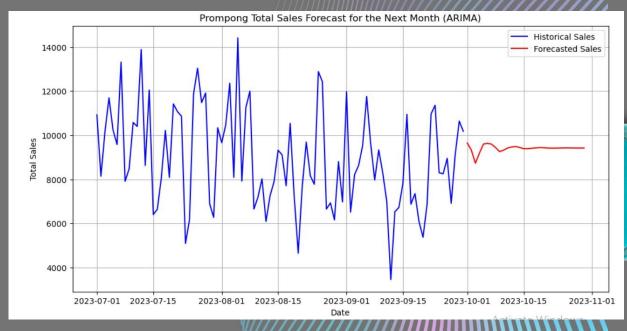


Prompong Branch Prediction for OCT 2023

Model: ARIMA
Time series dataset:
Total sales 3 month records
(July-Sep)

With only three months of historical data, We use an ARIMA model which are suitable for time series forecasting when having limited historical data and need to capture patterns and seasonality.

Although the accuracy of forecasts may improve with a larger dataset, but ARIMA can still provide valuable insight, predictions with the available information and display reasonable result.





Prompong Branch Prediction for OCT 2023

To evaluate the model, we imported NumPy and calculated the Mean of Actual Values from the real total sales data for October 1-10, then compared it with the Mean of Predicted Values from the ARIMA model.

The result shows the mean of the actual values is 9048.79 while the mean of the predicted values is approximately 9408.08
This suggests that, on average, the ARIMA model might be underestimating total sales.
However, this is a positive sign, indicating that real-world total sales are outperforming our expectations for Prompong branch.

actual_values (total sales
Oct 1-29/10 real world data) =
[10255, 8940, 12570, 9340, 7480,
11620, 13315, 9850, 8080,
8940,9470,9020,10600,7650,6940,
7330,8280,4460,8595,9690,11190,
8760,7560,7050,9140,8890,11790,
6480,9130]

The mean of actual values: 9048.793103448275

The mean of predicted values: 9408.08109127586

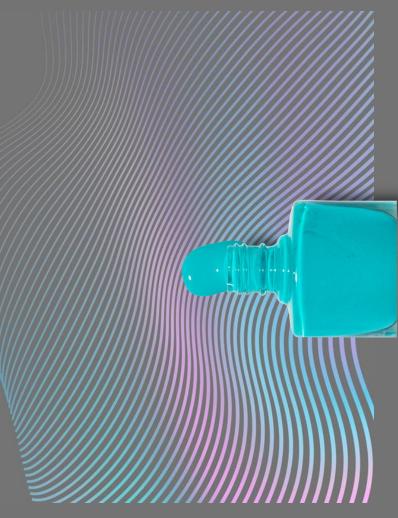
Total Sales	Forecast for the	Next	Month	(ARIMA):
	Forecasted Sales			
2023-10-01	9647.195509			
2023-10-02	9327.032299			
2023-10-03	8727.542250			
2023-10-04	9183.107159			
2023-10-05	9598.665224			
2023-10-06	9633.897366			
2023-10-07	9596.282620			
2023-10-08	9444.898005			
2023-10-09	9257.452422			
2023-10-10	9321.680142			
2023-10-11	9419.075583			
2023-10-12	9467.397358			
2023-10-13	9487.854398			
2023-10-14	9446.959159			
2023-10-15	9391.497854			
2023-10-16	9390.124064			
2023-10-17	9407.551492			
2023-10-18	9427.236493			
2023-10-19	9440.873962			
2023-10-20	9433.766527			
2023-10-21	9419.148602			
2023-10-22	9413.913756			
2023-10-23	9415.126426			
2023-10-24	9420.760672			
2023-10-25	9426.214946			
2023-10-26	9426.089080			
2023-10-27	9422.858697			
2023-10-28	9420.460618			
2023-10-29	9419.688964			
2023-10-30	9420.868087			

Prompong Branch Prediction for OCT 2023

To assess the accuracy of the model, we calculated the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics help us understand how closely the predicted values align with the actual values.

Mean Absolute Error (MAE): 1478.19 Mean Squared Error (MSE): 3926670.37 Root Mean Squared Error (RMSE): 1981.58

While lower MAE and RMSE values and a higher MSE are generally desirable, indicating better model performance, further model tuning or incorporating more historical data may lead to improved accuracy. This is particularly important in our case as we have limited time series data and need to consider real-world factors.



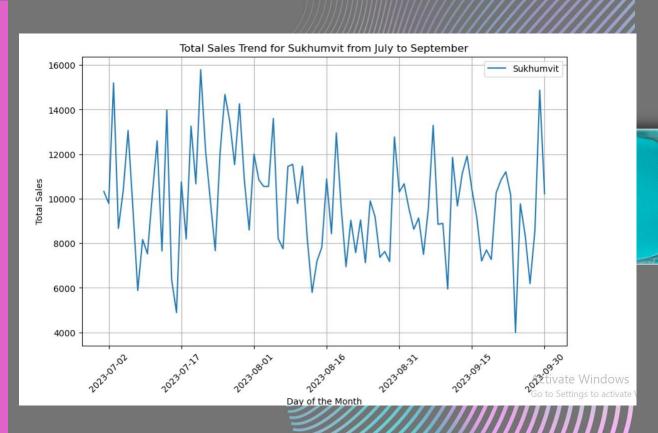
Sukhumvit Branch July-Sep

From July 2nd to July 17th, the sales volume ranged from approximately 5,000 to approximately 15,000 baht. There was a drastic drop in sales before July 17th, with values falling below 6,000 baht, followed by an increase to over 10,000 on July 17th.

After July 17th, total sales ranged between slightly under 8,000 but did not reach 16,000 baht, and then sales dropped until August 1st.

During the first half of August, the total sales had volumes between 6,000 and just under 14,000 baht. The graph dropped drastically before August 16th. After August 16th, sales remained between approximately 7,000 and 13,000 baht, increasing rapidly before August 31st

In September, from September 1st to September 15th, total sales had volumes between 6,000 and approximately 13,000 baht. There was a significant drop after September 15th, reaching a minimum sale of 4,000 baht on approximately September 20th before rising again, exceeding 14,000 baht before the end of September.

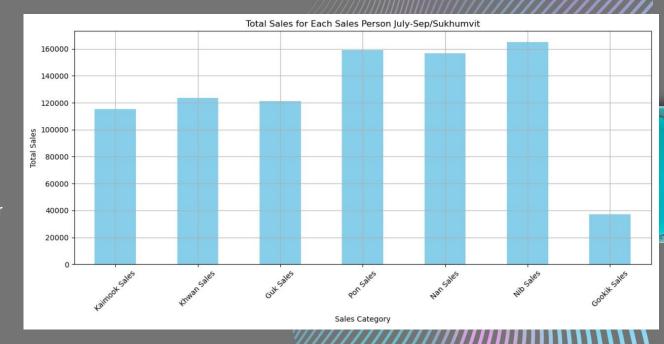


Sukhumvit Branch July-Sep

All employees' performance in terms of total sales falls within the range of over 100,000 to over 160,000 baht. Nib achieves the highest total sales. Three employees, Pon, Nan, and Nib, stand out with sales totaling over 140,000 baht each.

On the other hand, Kaimook, Khwan, and Guk have relatively close total sales, ranging from over 100,000 to slightly over 120,000 baht.

Gookik has room for improvement.

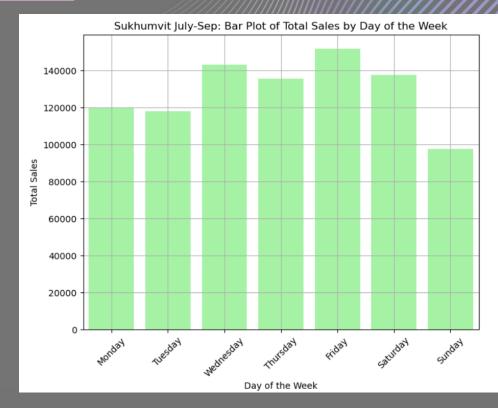




Sukhumvit Branch July-Sep

Sukhumvit: Bar Plot of Total Sales by Day of the Week

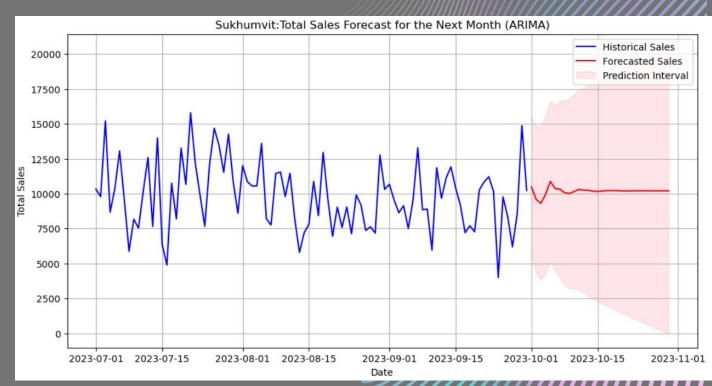
Display a bar plot comparing the days of the week across three months. Friday has the highest chance of achieving the highest sales, while Sunday has the lowest chance.





Sukhumvit Branch Prediction for OCT 2023 Sukhumvit:Total Sales Forecast fo

Model: ARIMA
Time series dataset:
Total sales 3 month
records (July-Sep)





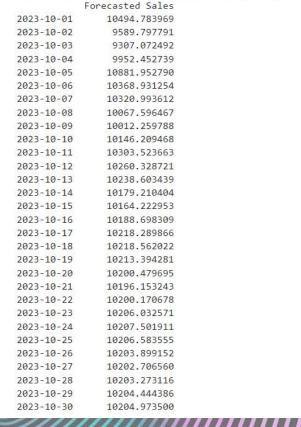
Sukhumvit Branch Prediction for OCT 2023

The difference between the Mean of Actual Values (8519.07) and the Mean of Predicted Values (approximately 10175.94) is not excessively large. This observation is made considering the short historical timeframe and other factors, such as economic conditions and a shortage of employees during the season.

Actual _values (total sales Oct 1-29/10 real world data) = [9770, 8980, 6760, 10940, 9740, 8100, 9280, 8685, 8945, 8850,7910,12120,6380,6990,7140,8150,8 335,9220,8280,10100,6660,3630,7770,62 50,9050,9245,12735]

Mean of Actual Values: 8519.074074074075

Mean of Predicted Values: 10175.941162703706



Total Sales Forecast for the Next Month (ARIMA):



Thonglor Branch July-Sep

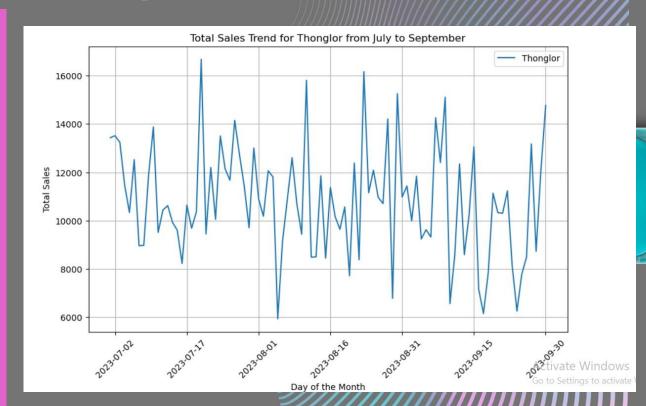
From July 2nd to July 17th, the sales volume ranged from over 8,000 to approximately 14,000.

After July 17th, the total sales for all 3 branches rose rapidly, exceeding 16,000 baht. Around July 20th, the sales dropped to under 10,000 baht but then stabilized at around 10,000 to 14,000 baht until the end of July.

During the first half of August, the total sales fluctuated significantly, experiencing both drops and rises. Before August 10th, it reached its lowest point at approximately 6,000 baht and then rose to just under 16,000 baht before August 16th.

From August 16th to 31st, the sales ranged from approximately 7,000 to just over 16,000 before dropping to under 12,000 by the end of the month.

From September 1st to September 15th, the sales volume ranged from approximately 7,000 to 15,000. After September 15th, the graph dropped to almost 6,000 baht before increasing and exceeding 14,000 baht by the end of September.

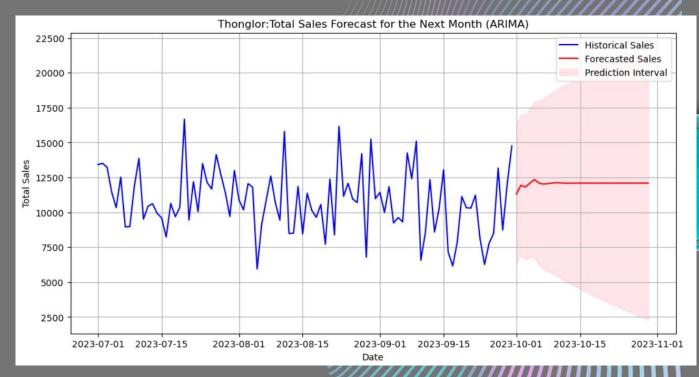




Thonglor Branch Prediction for

OCT 2023

Model: ARIMA Time series dataset: Total sales 3 month records (July-Sep)





Thonglor Branch Prediction for OCT 2023

With a Mean of Actual Values of approximately 9644.65 and a Mean of Predicted Values of approximately 12059.67

there is a notable difference between the actual and predicted values. This difference could be attributed to various factors, including the limited historical data available and external factors such as shortage of employees, which can significantly impact actual sales performance. Actual _values (total sales Oct 1-29/10 real world data) = [9460, 12940, 8160, 9795, 8120, 1290, 8010, 10165, 10960, 7430,12650,13085,13845,8340,61 30,7795,10845,10150,10090,1052 0,8005,9810,11160,10910,9710,1 4040,11005,7665,7610]

Mean of Actual Values: 9644.655172413793

Mean of Predicted Values: 12059.676401862067

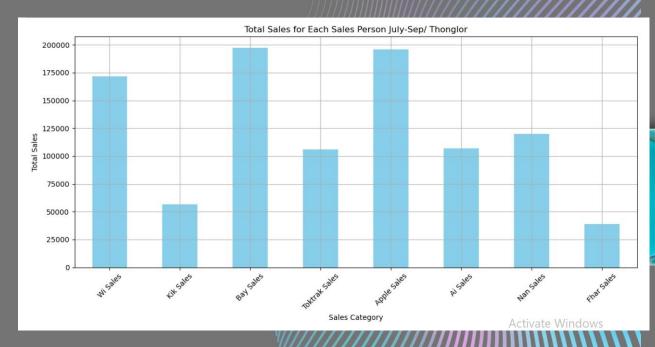
```
Thonglor: Total Sales Forecast for the Next Month (ARIMA):
            Forecasted Sales
2023-10-01
                11306.906278
2023-10-02
                11935.551412
2023-10-03
                11810.136008
2023-10-04
                12103.715808
2023-10-05
                12347.558009
2023-10-06
                12092.465176
2023-10-07
                12027.911149
2023-10-08
                12065.788598
2023-10-09
                12100.426125
2023-10-10
                12123.158339
2023-10-11
                12100.397544
2023-10-12
                12086.236316
2023-10-13
                12090.949970
2023-10-14
                12097.096381
2023-10-15
                12099.259320
2023-10-16
                12096.815997
2023-10-17
                12094.753686
2023-10-18
                12095.186800
2023-10-19
                12096.138094
2023-10-20
                12096.416257
2023-10-21
                12096.107442
2023-10-22
                12095.828533
2023-10-23
                12095.864114
2023-10-24
                12095.998541
2023-10-25
                12096.040565
2023-10-26
                12095,999960
2023-10-27
                12095.961909
2023-10-28
                12095.964539
2023-10-29
                12095.982784
2023-10-30
                12095.989270
```



Thonglor Branch July-Sep

The most outstanding employees are Wi, Bay, and Apple, with Bay and Apple achieving total sales exceeding 175,000 and almost reaching 200,000 baht over the course of three months.

In a branch with 8 staff members, there's a significant gap in total sales between these top three performers and the other 5 individuals. This indicates a need for improvement and emphasizes the importance of teamwork among the entire staff, including the top three performers.

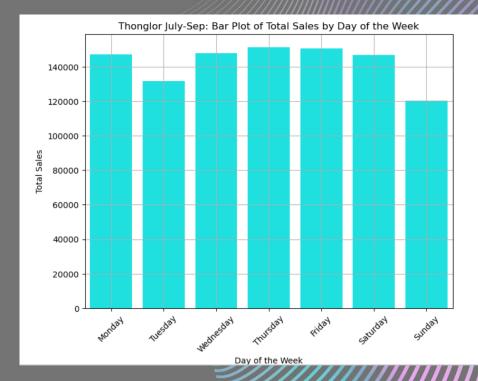




Thonglor Branch July-Sep

Thonglor: Bar Plot of Total Sales by Day of the Week

Display a bar plot comparing the days of the week across three months. Thursday and Friday have the highest chance of achieving the highest sales equally, while Sunday has the lowest chance.





Predict Total Sales Across All Data Frames

Use Random Forest Regression Model

The model was evaluated by splitting the data into a training set and a testing set to assess its accuracy

Average Mean Absolute Error (MAE): 179.47
The MAE measures the average absolute difference between the actual and predicted sales.

An MAE of 179.47

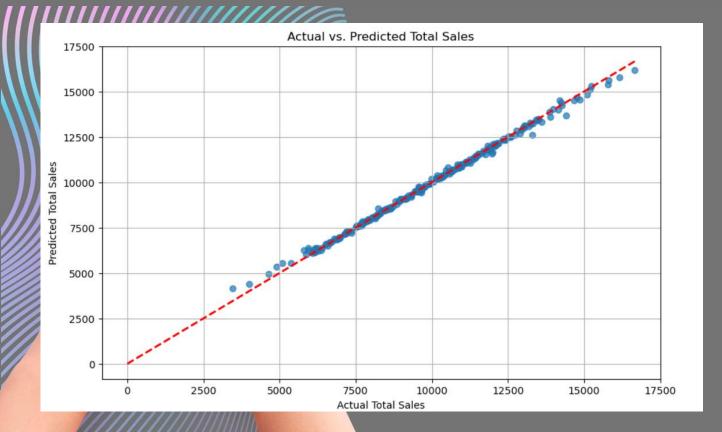
suggests that, on average, the model's predictions are off by this amount in terms of total sales.

Average Mean Squared Error (MSE): 135473.79The MSE calculates the average of the squared differences between actual and predicted values. The value of 135473.79 is relatively low, indicating that the model's predictions are, on average, close to the actual values.

Average R-squared (R^2): 0.96:. An R-squared of 0.96 is quite high and indicates that the model explains a substantial amount of the variance in total sales.







Predict Total Sales Across All Data Frames

Total Sales across all data frames: 2,724,185.00: This is the total actual sales across all three nail salon branches.

Total Predicted Total Sales Across All Data Frames: 2,723,521.00: This is the total predicted sales across all three branches.

Model Interpretation:

The relatively small difference between the total actual sales and the total predicted sales, with only a 664.00 difference, is a positive sign.

It suggests that the model is performing well in capturing the overall trend in sales for the three branches.

While the model may not provide an exact match,

the small difference indicates that it's quite accurate in its predictions.

Business Interpretation: Total Predicted Sales Across All Data Frames (2,723,521.00) is lower than the Total Actual Sales Across All Data Frames (2,724,185.00), it indicates a slightly lower prediction compared to the actual sales. In this context, it could be considered a less positive sign for the business as the model slightly underestimated the total sales.

However, it's important to consider other factors and trends in the data and real world conditions to make a more comprehensive assessment.



Summary of Result

Sales Trends Across Three Nail Salon Branches:

Over the three months (July-September 2023), all three nail salon branches experienced fluctuations in their total sales.

The **Thonglor** branch consistently had the highest total sales, followed by **Sukhumvit** and **Prompong**.

Total sales across all three branches showed a declining trend over this period.

Employee Performance:

Employee performance significantly impacted total sales. Some employees consistently achieved high sales, while others performed below average.

Aom, Nib, Pon, and Nan were among the top-performing employees, contributing significantly to total sales.

Impact of Historical Data:

With a limited three-month historical dataset, ARIMA and Random Forest Regression models were used for predicting future total sales.

The ARIMA model provided predictions that were slightly lower than the actual sales, indicating a conservative outlook.



Summary of Result

Predictive Models:

The ARIMA model provided a reasonably conservative estimate of future sales trends with the available data.

The Random Forest Regression model showed promise in predicting total sales, though further validation and refinement may be necessary.

Recommendations:

The analysis suggests room for improvement in employee performance, especially among staff with lower sales figures.

Strategies such as staff training, sales incentives, and promotions could be implemented to enhance total sales.

More historical data would likely improve the accuracy of sales predictions.



conclusion

Analysis and Insights from Provided Data:

The analysis of historical sales data for multiple nail salon branches (Prompong, Sukhumvit, Thonglor) has revealed valuable insights into sales trends and patterns.

The use of time series analysis, specifically the ARIMA model, has shown potential for predicting future sales, with the model's performance varying based on available historical data and real-world factors.

Evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) have been used to assess the accuracy of predictive models. These metrics provide quantitative measures of model performance.

** Real-world factors, including variations in employee performance, marketing efforts, and external market conditions, play a significant role in influencing sales outcomes

Problems/Solutions

How can we accurately predict the total sales for each of the three nail salon branches? ARIMA and Random Forest Regression can be used to accurately predict the total sales for each nail salon branch.

For example, in the analysis, ARIMA was applied to historical sales data from 3 branches, providing precise forecasts for the upcoming month. By examining the historical data, ARIMA identified patterns and trends specific to each branch and used them to make predictions, allowing for branch-level sales estimations.

Can we reliably estimate the likelihood of various sales outcomes for the nail salon branches? In the analysis, ARIMA and Random Forest Regression were used to predict sales outcomes, and they provided probability distributions for different sales levels.

By assessing these probabilities, we can estimate the likelihood of achieving various sales outcomes, enabling a data-driven approach to decision-making.

How can predictive models be effectively leveraged to reasonably estimate future sales and assist in data-driven decision-making for the nail salon branches?

For instance, in the analysis, ARIMA forecasted a decrease in total sales for a specific branch in the upcoming month based on historical data. This information allowed branch managers to make data-driven decisions, such as adjusting staffing levels and planning targeted promotions, to address the anticipated decline.

Appendix

Appendix: Relevant Project Assets

1. Data Sets:

Historical sales data for each nail salon branch (July-September 2023)

Employee performance data, including individual sales data for each employee

2. Code Repositories:

GitHub repository with the Python scripts for data preprocessing and modeling

3. Visualizations:

Line graphs illustrating the sales trends for each nail salon branch (July-September 2023) Graphs depicting employee performance, showcasing individual sales figures

4. Model Summaries:

Summaries of ARIMA and Random Forest Regression models used for sales predictions Evaluation metrics and performance results for each model

5. Detailed Analysis:

Detailed reports on sales trends and patterns identified during the analysis
In-depth statistical analyses, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) calculations

