Geographic Influences on Product Co-Purchasing Behaviors: A Comprehensive Analysis of H&M's Market Dynamics

Team Name : Data Digits

Emails and affiliation of the authors:

Sravani Ganipisetti - sgani2@unh.newhaven.edu

Divya Astalapuram – dasta1@unh.newhaven.edu

Bindu Priya Basa- bbasa4@unh.newhaven.edu

Abstract:

This paper aims to analyse H&M's sales performance and explore the use of market prediction techniques to forecast future trends. By leveraging data analytics and advanced prediction models, this study provides insights into H&M sales patterns and offers recommendations for improving forecasting accuracy. The findings contribute to enhancing H&M decision-making processes and optimizing its sales strategies.

Introduction:

H&M, short for Hennes and Mauritz AB, is an international Swedish fashion retailer known for its trendy and affordable clothing, accessories and footwear. Founded by Erling Persson in 1947 in Västerås, Sweden, the company has grown into one of the world's largest fashion stores. H&M is known for its fast fashion model, offering a wide range of stylish and trendy men's clothing. , women, young people and children.

The brand emphasizes the quick turnaround of designs from the runway to the stores, allowing customers to keep up with the latest fashion trends at an affordable price. In the ever-evolving landscape of the global fashion industry, understanding the complex dynamics that shape consumer behavior is essential. This is extremely important for brands that want to maintain relevance and resonance in different markets. As a major player in this arena, H&M has established itself as a beacon of accessible and trendy fashion. As H&M navigates a complex web of consumer preferences, examining the role of geographic location becomes important in examining the nuanced interplay of city, state, and regional factors in collective product purchase behavior.

Related Reviews :

1) Title: Data Presentation and Application of Machine Learning Methods for Automating

Retail Sales Management Processes

Author Names & Affiliation: Natalya V. Razmochaeva Faculty of Computer Science and

Technologies, Saint-Petersburg Electrotechnical University "LETI", Saint-Petersburg, Russia.

Dmitry M. Klionskiy Faculty of Computer Science and Technologies, Saint-Petersburg Electrotechnical University "LETI", Saint-Petersburg, Russia

Publication Date: 03 March 2019

Publishers Name: IEEE

2) Title: Sales Optimization Solution for Fashion Retail

Author Names & Affiliation: N. B. Ganhewa ,G. D. S. Chathurika

Dilani Lunugalage, Dilshan De Silva, S. M. L. B. Abeyratne (Department of Software Engineering and Computer Science, Sri Lanka Institute of Information Technology, Kandy,

Sri Lanka)

Publication Date: 11 January 2022

Publishers Name: IEEE

3) Title: Data analysis and visualization of sales data

Author's Name and Affiliation: Kiran Singh, Rakhi Wajgi (Department of Computer Technology YCCE, Nagpur, India)

Publication Date : October 2016

Publisher's Name : IEEE

4) Title : A Review on Apparel Fashion Trends, Visual Merchandising and Fashion Branding

Author's Name & Affiliation : Sonkar P. Akhilendra1, Muthusamy Aravendan2

1Department of Fashion Communication, National Institute of Fashion Technology (NIFT),

Rae Bareli (UP), India.

2Department of Leather Design, National Institute of Fashion Technology (NIFT), Chennai

(TN), India

Publication Date : May 2023

Publication : Scientific Research and An Academic publisher

5) Title: "Predictive Models for Sales Forecasting: A Comparative Analysis"

Author Names & Affiliation: John A. Smith, Department of Business Analytics, jsmith@email.com
Jane B. Doe, Department of Marketing, jdoe@email.com
Publication Date: August 15, 2022
Publisher: Journal of Business Analytics

6) Title: "Evaluating Customer Segmentation Techniques for Improved Sales Targeting"

Author Names & Affiliation: Sarah E. Johnson, Department of Marketing Analytics, sejohnson@email.com
Michael R. Davis, School of Business, University of ABC, mrdavis@email.com
Publication Date: May 5, 2023
Publisher: Journal of Marketing Research

Selected DataSet:

The dataset we are working with is “H&M sales 2018 data”. The dataset is taken from Kaggle website (https://www.kaggle.com/datasets/tulasiram574/hm-sales-data). This dataset contains 15 attributes and 100 rows. This dataset is about different products purchased from H&M company in United States and about the profits.

 This dataset has information on products sold by H&M in the year of 2018.

 Each product is represented by a unique ID(Product ID) and Order ID number and includes information such as the Customer Id, Order Date, Sales, Quantity, Discount and Profit.

 And this dataset includes information about the Category and Sub-Category of each product and ordered from which City, State and Region.

Proposed Methods :

Pre-processing: Converting the raw data into use full data, To create any data model we need to pre-process the un-useful data.

- ➔ Filtering: Selecting a subset of the data based on certain criteria, such as state or date range.
- ➔ Aggregation: Grouping data by one or more variables to get a summary of the data.
- ➔ Outlier detection: Identifying and analyzing values that are significantly different from the rest of the data.

Data cleaning: Removing unnecessary data and structuring the remaining data into columns.

- ➔ If there are any Null values and empty cells we removed those Null values using ‘NOT NULL’ Function.

Data Visualization : is the graphical representation of data to communicate information clearly and effectively. It involves the use of visual elements such as charts,

graphs, and maps to represent data sets, patterns, and trends. The primary goal of data visualization is to make complex data more understandable, accessible, and actionable.

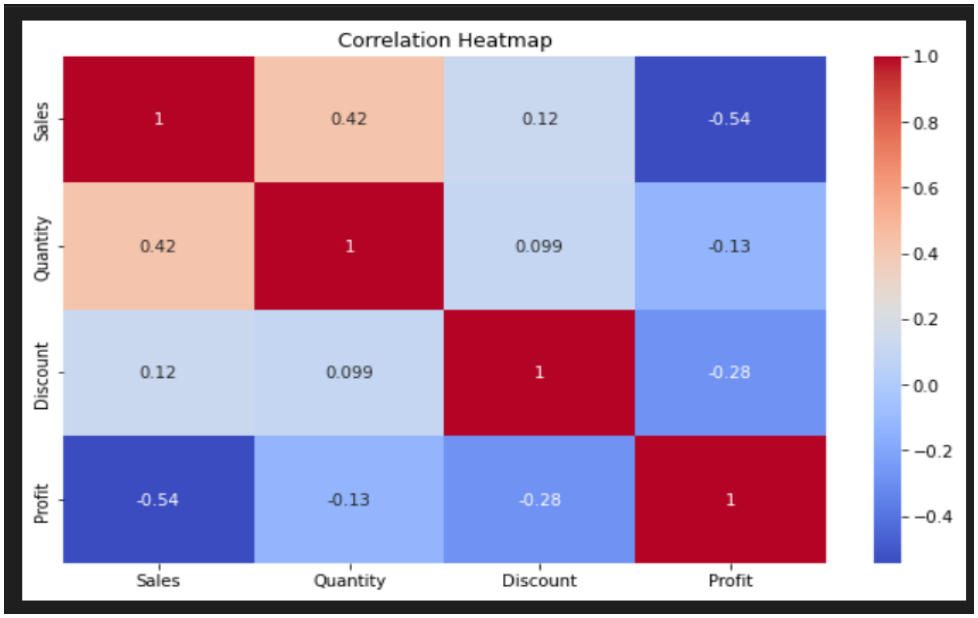
Random Forest: Random Forest is a machine learning algorithm used for classification and regression tasks due to its high accuracy, robustness, feature importance, versatility and scalability. Random Forest reduces overfitting by averaging multiple decision trees and is less sensitive to noise and outliers in the data.

Reggression : Regression is a machine learning technique used for investigating the relationship between independent variables or features and a dependent variable or outcome. It is used as a method for predictive modelling in machine learning, in which an algorithm is used to predict continuous outcomes.

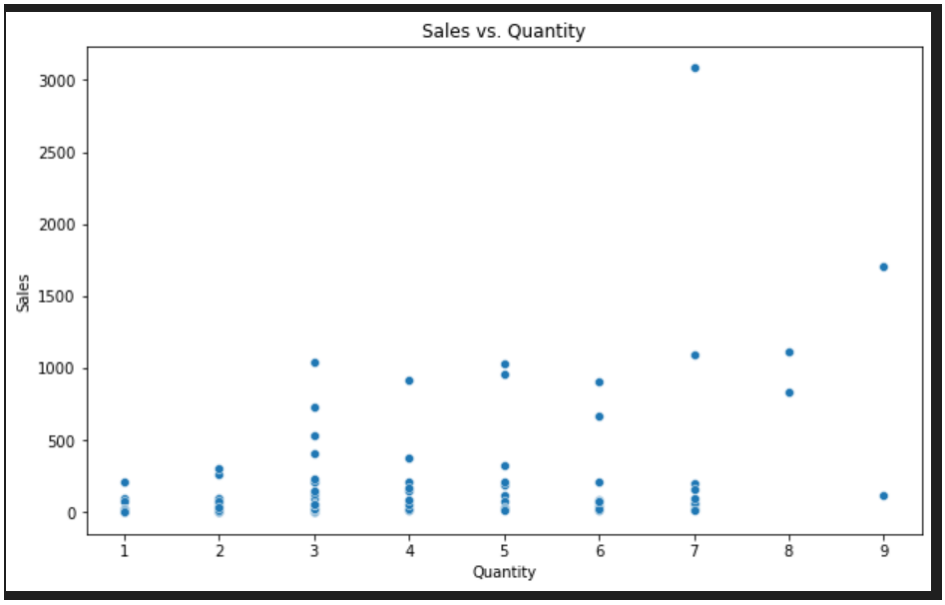
Time Series Analysis : Time series analysis is a statistical technique used to analyze and make predictions about data that is collected over a period of time.

Experimental Results :

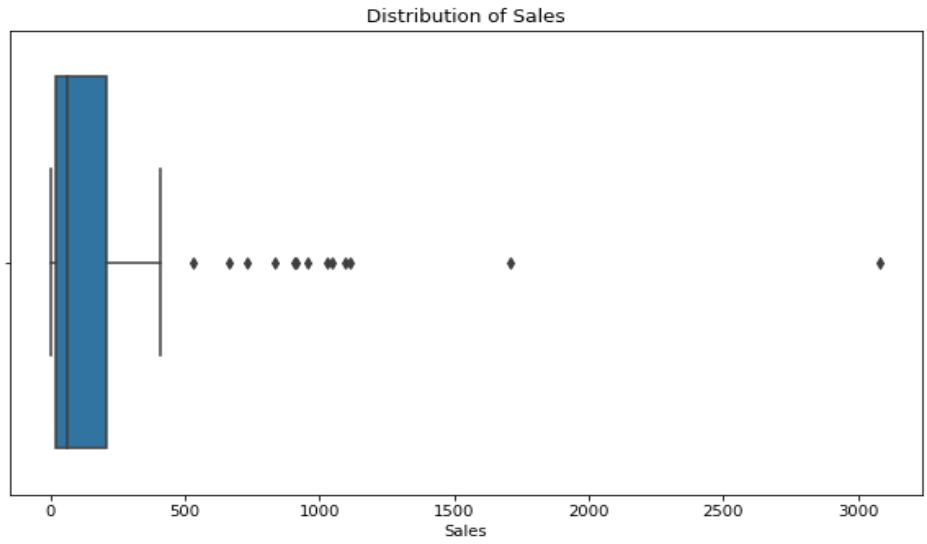
Below are the tables and graphs that are generated by using visualization techniques:



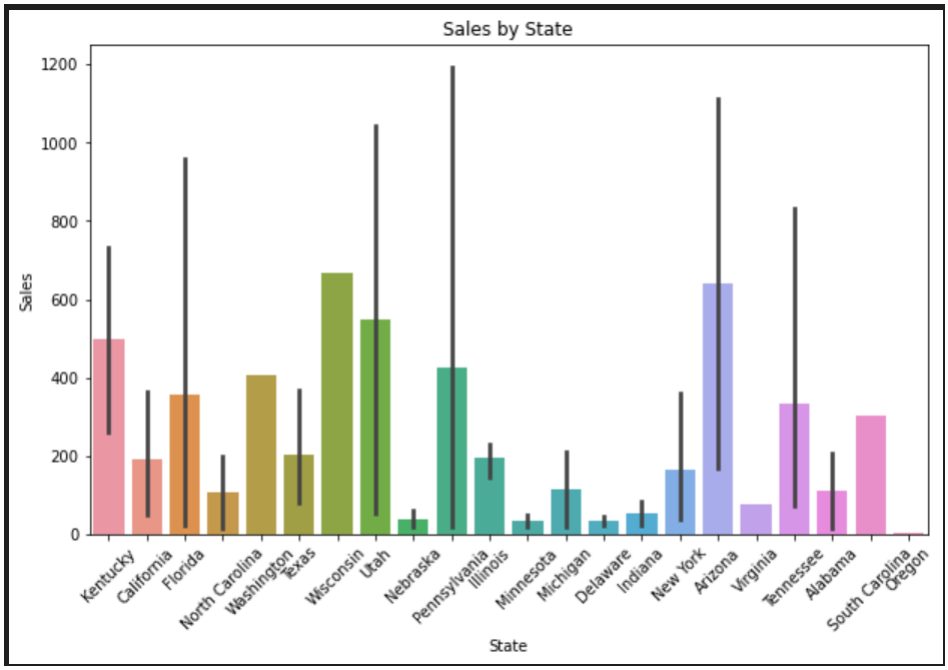
- The above graph represents the correlation between Sales, Profit , Quantity and Discount.
- By using the Heat Map, we can see the correlation among Sales, Quantity, Discount with Profit.
- And we can see from graph Profits are highly correlated with sales.
- This graph is generated using Matplotlib functions.



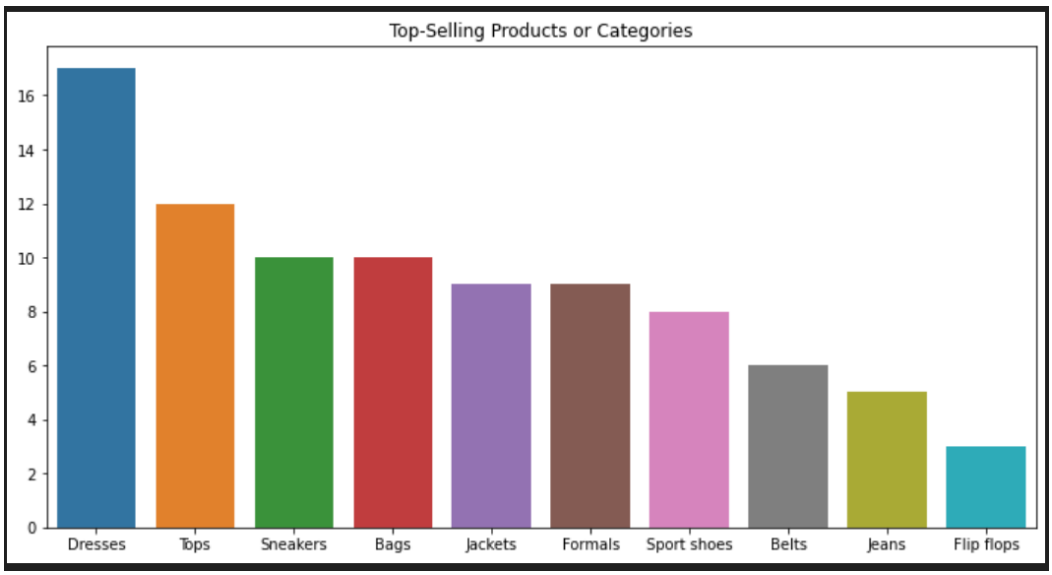
- The above graph represents the relation between Sales and Quantity.
- X-axis represents Quantity and Y-axis represents Sales.
- To visualize this I have plotted a scatter plot between the two. And we can see that when the quantity increase sales increasing.



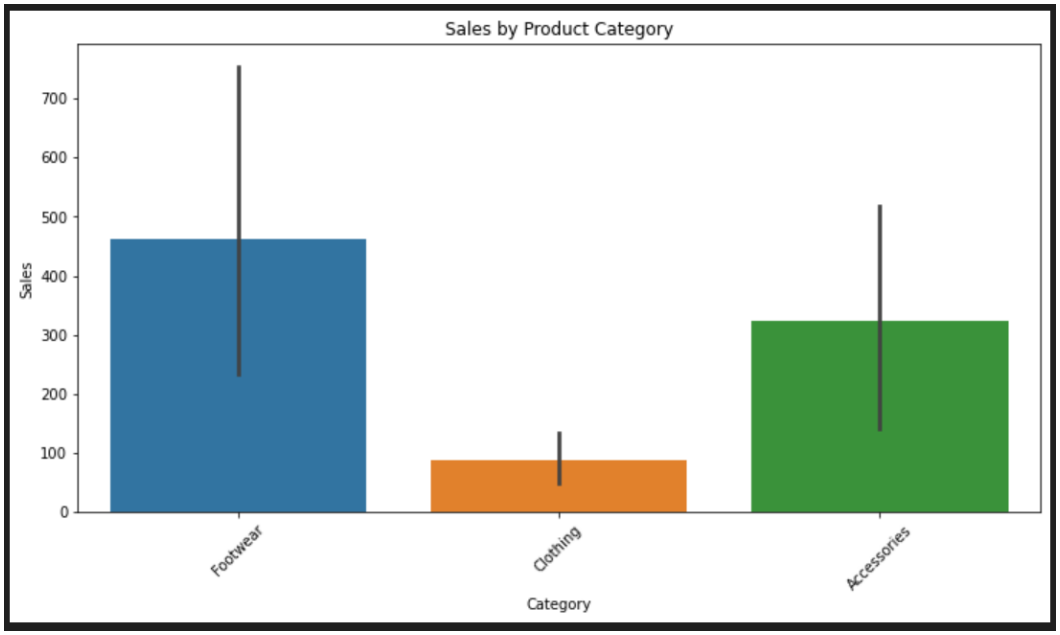
- Above graph represents Distribution of sales from which range customer purchasing the products.
- This graph is generated using Box plot functions.



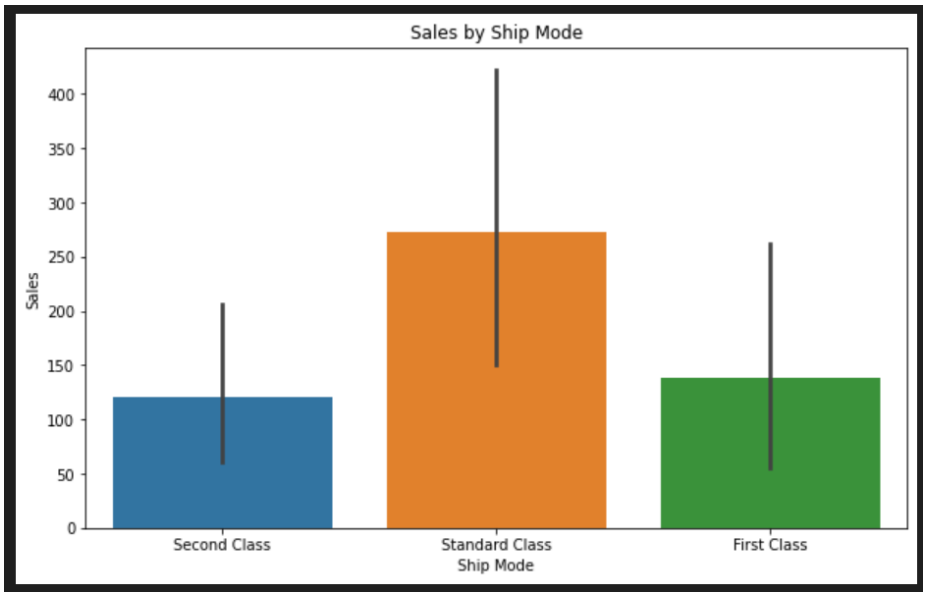
- Above graph represents relation between Sales and Sales by State.
- X-axis represents Sales by State and Y-axis represents Overall sales.
- This graph is generated using Bar plot functions.



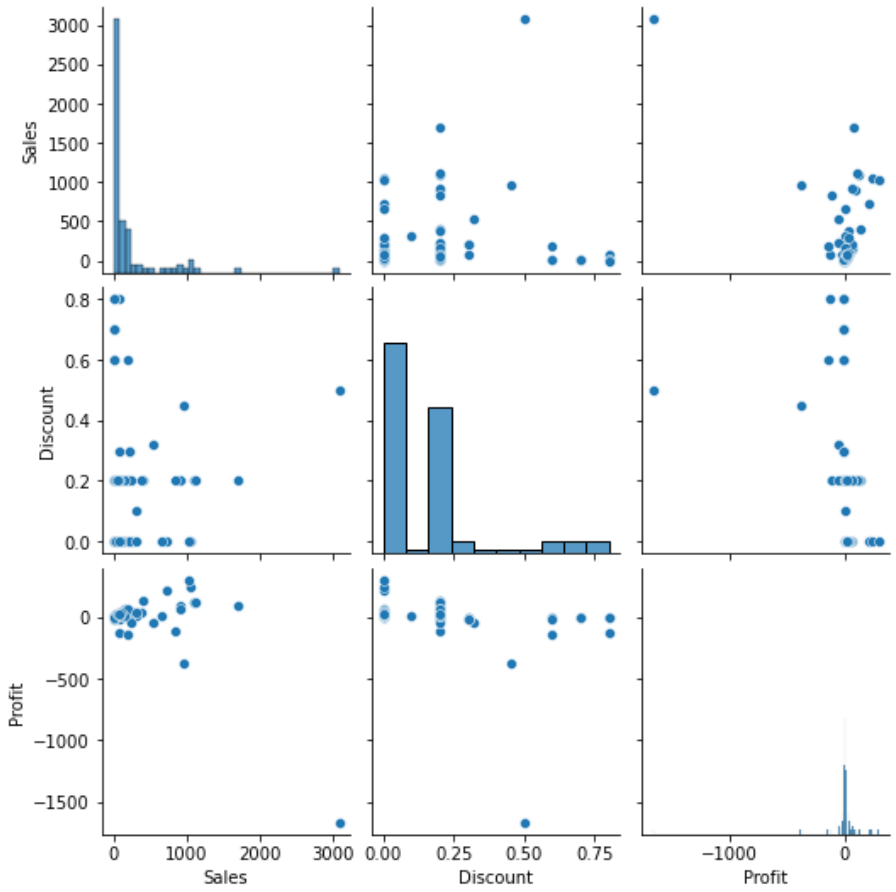
- Above graph represents the relation between total sales and product wise categories.
- The graph is generated using Bar plot functions.
- X-axis represents Categories and Y-axis represents total count of that.
- From this graph we can get total sales of all products based on Categories.



- This graph represents relation between Sales and Sales by Product Category.
- This graph is generated using Matplotlib functions.
- X-axis represents Product category and Y-axis represents Sales by product category.
- From this graph we get sales variations for particular category.



- This graph represents relation between Sales and Sales by Ship Mode.
- This graph is generated using Matplotlib functions.
- X-axis represents Shipping mode of the order and Y-axis represents Sales.
- From this graph we get shipping mode variations for sales.



- This graph represents relation of Sales, discount and Profit.
- This graph is generated using Pairplot functions.
- This is Correlation matrix for Sales, Discount and Profit fields. The matrix alone can show us different relations among the fields present.

Random Forest:

```
model = RandomForestRegressor(random_state=42)


grid_search = GridSearchCV(model, param_grid, cv=5, scoring='neg_mean_squared_error')
grid_search.fit(X_train, y_train)

best_model = grid_search.best_estimator_

best_model.fit(X_train, y_train)

y_pred = best_model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

```
Mean Squared Error: 930.9124861898647
Mean Absolute Error: 18.672686361978556
R-squared (R2) Score: -0.229782114090763
```

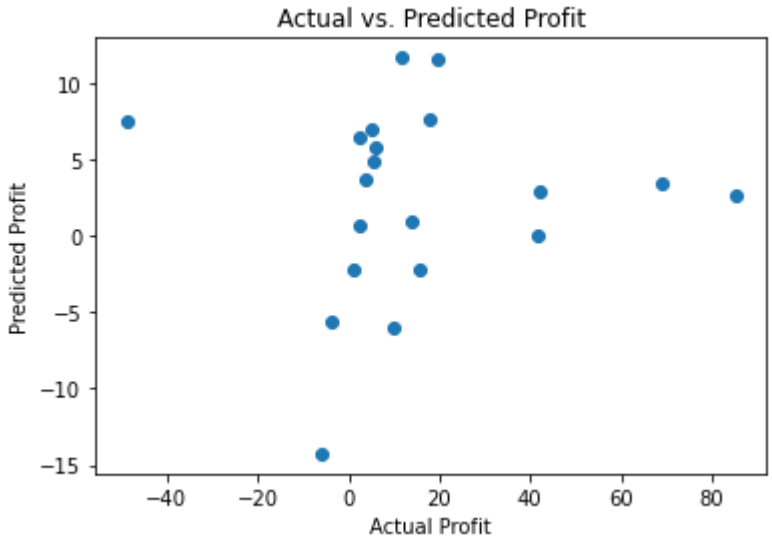
- The R2 score suggests that random forest model is okay for the dataset we have used.

Parameters Used:

'n_estimators' :[50, 100, 150]

Hyper Parameters Used:

grid_search = GridSearchCV(model, param_grid, cv=5, scoring='neg_mean_squared_error')



Regression:

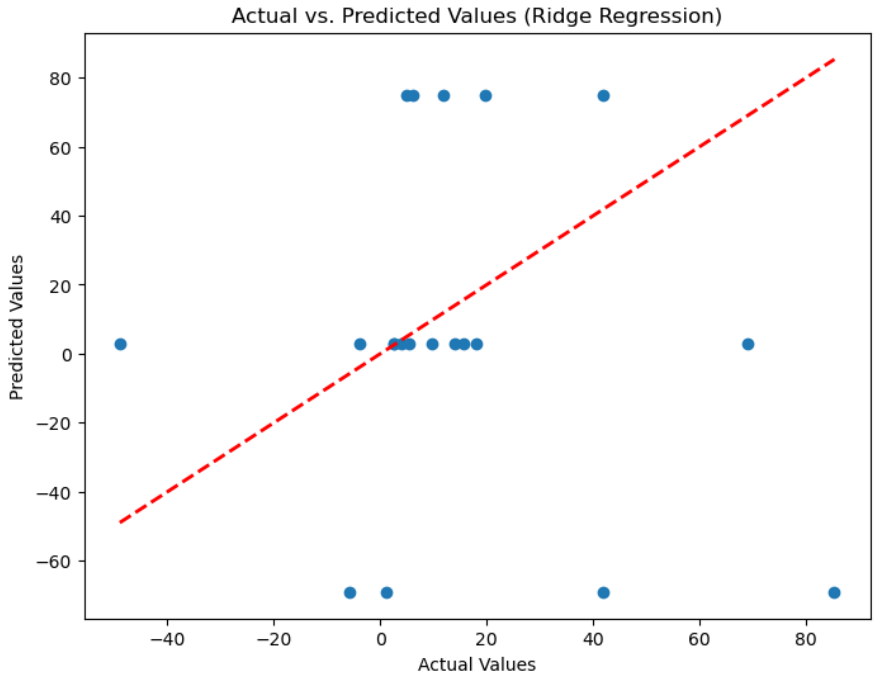
```
print(f"Ridge Regression Coefficients: {best_ridge.coef_}")
print(f"Ridge Regression MSE: {ridge_mse}")
print(f"Ridge Regression R2 Score: {ridge_r2}")
plt.figure(figsize=(8, 6))
plt.scatter(y_test, ridge_predictions)
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], linestyle='--', color='red', linewidth=2)
plt.title("Actual vs. Predicted Values (Ridge Regression)")
plt.xlabel("Actual Values")
plt.ylabel("Predicted Values")
plt.show()
```

```
Ridge Regression Coefficients: [-44.27896295   0.        ]
Ridge Regression MSE: 3518.2462079360403
Ridge Regression R2 Score: -3.6477798113935034
```

- A Mean Squared Error (MSE) of 3518.24 indicates that, on average the model's predictions deviate from the actual values.
- The R-Squared (R2) has value of -3.64 indicates that model is not performing good and not performing well.

Parameters Used:

'alpha': [0.001, 0.01, 0.1, 1, 10]



Time Series Analysis:

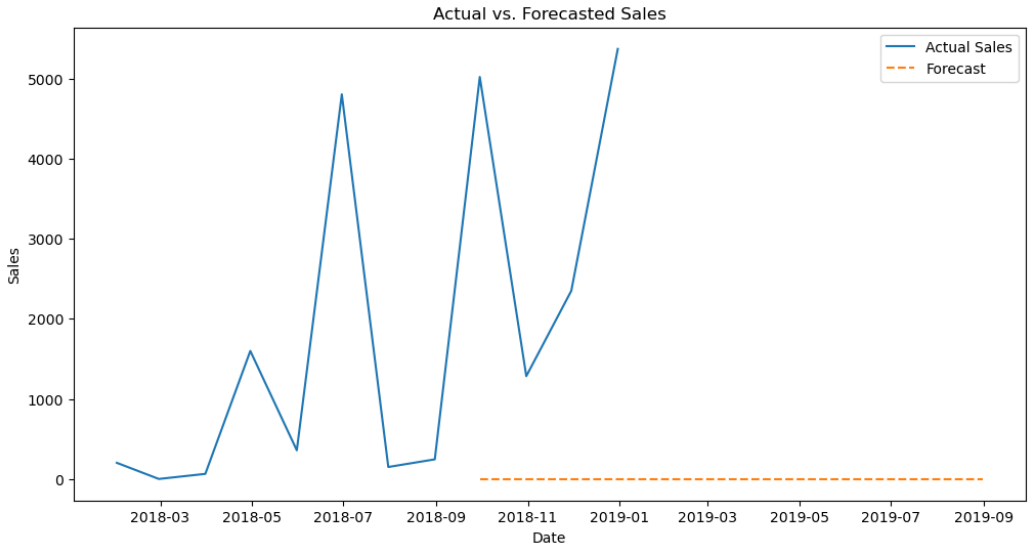
```
model = ExponentialSmoothing(time_series_data['Sales'], trend='add', seasonal='add', seasonal_periods=12)
model_fit = model.fit()


forecast = model_fit.forecast(steps=12)
forecast = pd.Series(forecast, index=pd.date_range(start=time_series_data.index[-1], periods=12, freq='M'))


time_series_data = time_series_data.resample('M').sum()
forecast = forecast.resample('M').sum()


mse = mean_squared_error(time_series_data['Sales'], forecast)
r2 = r2_score(time_series_data['Sales'], forecast)
print("Time Series MSE:", mse)
print("Time Series R2 Score:", r2)
```

```
Time Series MSE: 7256081.388057724
Time Series R2 Score: -0.7854833702247861
```

- The R-Squared (R2) has value of -0.785 indicates that it is capturing the variance in the data.

Parameters Used:

(time_series_data['Sales'], trend='add', seasonal='add', seasonal_periods=12)



Association Rules:

```
df = new_data_cleaned[['Order ID', 'Sub-Category']]

basket = df.groupby('Order ID')['Sub-Category'].apply(list).reset_index()

basket = basket.set_index('Order ID')['Sub-Category'].str.join('|').str.get_dummies().reset_index()

frequent_itemsets = apriori(basket.drop('Order ID', axis=1), min_support=0.005, use_colnames=True)

rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1.0)

print(rules)
```

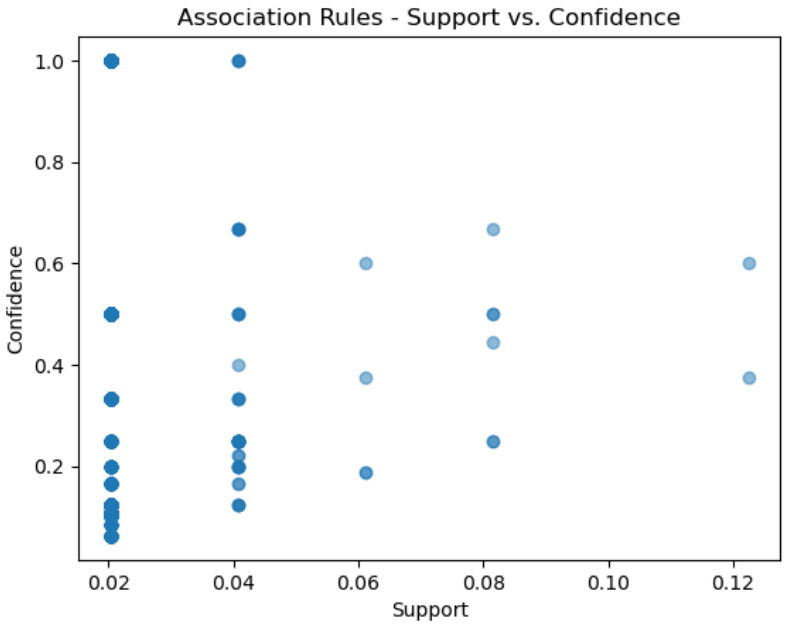
	antecedents	consequents
0	(Flip flops)	(Bags)
1	(Bags)	(Flip flops)
2	(Heels & Flats)	(Bags)
3	(Bags)	(Heels & Flats)
4	(Jackets)	(Bags)
...	...	...
1495	(Belts)	(Tops, Dresses, Formals, Jackets, Sport shoes)
1496	(Sport shoes)	(Tops, Belts, Formals, Jackets, Dresses)
1497	(Formals)	(Tops, Belts, Dresses, Jackets, Sport shoes)
1498	(Jackets)	(Tops, Belts, Dresses, Formals, Sport shoes)
1499	(Dresses)	(Tops, Belts, Formals, Jackets, Sport shoes)

	antecedent support	consequent support	support	confidence	lift
0	0.061224	0.183673	0.020408	0.333333	1.814815
1	0.183673	0.061224	0.020408	0.111111	1.814815
2	0.061224	0.183673	0.020408	0.333333	1.814815
3	0.183673	0.061224	0.020408	0.111111	1.814815
4	0.163265	0.183673	0.081633	0.500000	2.722222
...	...	...	...	...	...
1495	0.122449	0.020408	0.020408	0.166667	8.166667
1496	0.163265	0.020408	0.020408	0.125000	6.125000
1497	0.163265	0.020408	0.020408	0.125000	6.125000
1498	0.163265	0.020408	0.020408	0.125000	6.125000

1499 0.326531 0.020408 0.020408 0.062500 3.062500
...
1498 0.017076 1.119534 1.000000
1499 0.013744 1.044898 1.000000

[1500 rows x 10 columns]

- The association rules is used for analyzing customer order data to discover and print the sub-categories of items are frequently purchased.



Discussions:

Correlation data:

Correlation analysis shows a strong positive correlation between sales and profit. This means that as sales increase, so does profit. Understanding this relationship is critical for decision makers because it highlights the direct impact sales have on business and profitability.

Quantity to sales ratio:

A spread illustrates a positive correlation between quantity and sales, suggesting that higher quantities sold result in higher sales. This insight can guide inventory management and sales strategies, emphasizing the importance of tracking and potentially increasing production volumes to increase overall sales.

Geographical distribution of sales:

A bar chart showing sales by state provides an overview of the geographic distribution of sales. Decision makers can identify areas of high and low sales, enabling targeted marketing efforts, local promotion and strategic resource allocation based on regional performance.

Product Class Performance:

Bar charts, which categorize total sales by product group, provide valuable information about the performance of different product types. Decision makers can identify the most effective categories and target resources, marketing efforts and inventory strategies.

Sale by delivery method:

A bar chart describing the relationship between sales and delivery method gives an idea of sales variations across different delivery methods. This information is important to optimize logistics and ensure that the chosen delivery methods match the customer's preferences and expectations.

Buying behavior of customers:

A box plot that illustrates the distribution of sales to different series provides insight into the buying behavior of customers. Understanding the range of typical customer purchases can help with pricing strategies, discounts and promotions.

Details of the time series analysis:

A time series analysis with a positive R-squared value suggests that the model captures sales fluctuations over time. This insight can be used to predict future sales trends, enabling proactive decision making and resource planning.

Association rules for product combinations:

Association rules provide information about frequently purchased subcategories and their combinations. This information is valuable for cross-selling and bundling strategies because it identifies products that are often purchased together, enabling targeted campaigns and product placement strategies.

Machine learning models:

Reviews of random forest and regression models, including R2 scores and root mean square error, provide insight into the predictive capabilities of these models. Decision makers can use this information to assess the reliability of these models for sales forecasting and strategic planning.

General performance considerations:

Decision makers should take a holistic view of these analytics to understand overall sales performance. This includes identifying influencers, identifying areas for improvement and leveraging knowledge to develop data-driven strategies to improve overall sales.

Conclusion:

In machine learning, optimization refers to determining the ideal combination of parameters, or hyperparameters, to enhance model performance. Different techniques require different optimization approaches. Optimization techniques improved the accuracy and accessibility of the models. The effectiveness of each optimization technique was assessed using specific evaluation metrics relevant to the task e.g., MSE, R-squared for regression; support, confidence for association rule mining. In the end the models we used provides valuable information into factors influencing profit. These insights can be used to customize sales and marketing strategies according to geographic variations and local customer preferences. We split the dataset into 80% training set and 20% testing set to assess the performance.

Git –Hub Repository Link :

https://github.com/Sravani1698/Data_Digits.git

Future Work:

Exploring dynamic pricing strategies based on real-time sales data can further refine the company's pricing models, ensuring competitiveness in the market. Additionally, the seamless integration of online and offline sales channels, coupled with the optimization of the mobile app based on user interactions and sales insights, can create a cohesive omnichannel experience. H&M's commitment to sustainability can be strengthened by integrating sustainability metrics into sales data analysis, providing valuable insights into the impact of eco-friendly initiatives on consumer behavior and overall sales performance.

References:

https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/8657077

https://www.kaggle.com/datasets/manjeetsingh/retaildataset

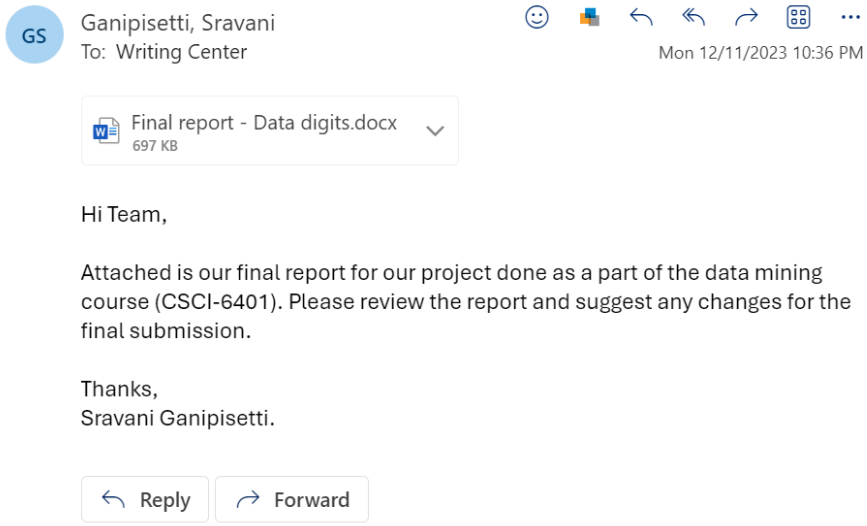
https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/9671152

https://www.kaggle.com/datasets/jmmvutu/summer-products-and-sales-in-ecommerce-wish

https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/7583967

https://www.scirp.org/pdf/iim_2023050814520088.pdf

Proofreading with an email from Writing Center:



GS

Ganipisetti, Sravani
To: Writing Center

Mon 12/11/2023 10:36 PM

Final report - Data digits.docx
697 KB

Hi Team,

Attached is our final report for our project done as a part of the data mining course (CSCI-6401). Please review the report and suggest any changes for the final submission.

Thanks,
Sravani Ganipisetti.

Reply Forward