**Team Project  
  
  
Fraud Detection in E-Commerce Transactions: Analyzing Patterns and Anomalies in Online Retail Data**

## **Data Diggers**

## **Provide Team Members’ Names and UIS Emails**

|  |  |
| --- | --- |
| **Names** | **UIS Email ID** |
| Vinay Kumar Kolukula Pally | vkolu@uis.edu |
| Shiva Boda | Sboda@uis.edu |
| Charitha Banda | cband4@uis.edu |
| Deepthi Tamma | dtamm2@uis.edu |

**Abstract**

This task looks at an enormous dataset of online retail exchanges to recognize client buying designs, conjecture deal volumes, and create noteworthy business experiences. The review will utilize a mix of exploratory and prescient investigation to see better the connection between item costs, client conduct, and geographic patterns.

The exploratory examination uncovered huge client fragments utilizing K-implies and various leveled grouping strategies. These fragments were characterized in view of normal unit cost, exchange volume, and buying conduct across nations. The discoveries uncovered unmistakable geographic and item drifts, like territorial varieties in client inclinations and explicit items that drive higher deal volumes.

Prescient examination utilized relapse models, choice trees, and brain organizations to conjecture exchange amounts. Models were assessed on key execution measurements, for example, RMSE and R-squared, with the relapse model exhibiting high prescient exactness. These forecasts give bits of knowledge into future interest designs, empowering organizations to further develop stock administration and asset assignment.

The concentrate additionally utilized representations, including country-wise normal unit cost and amount dispersions, to improve interpretability. Furthermore, time-series patterns offered a nitty gritty perspective on buying conduct over the long run. The significant proposals obtained from this examination incorporate refining estimating systems, improving designated showcasing endeavors, and enhancing the inventory network the executives to fulfill client needs.

This undertaking gives an information-driven system to further developing dynamics in the web-based retail area, adding to expanded effectiveness, consumer loyalty, and income development.  
  
**Problem Description:**

Background and Importance

The internet-based retail industry is recognized by the huge and complex exchange of information, making it challenging to separate significant bits of knowledge and upgrade business systems. Understanding client buying conduct, deal patterns, and key income drivers is basic to staying serious in this industry. This venture means to resolve these issues by utilizing progressed information examination to find stowed-away examples, actually fragment clients, and figure exchange results.

The dataset utilized in this study contains conditional information from a worldwide web-based business organization, including item unit cost, amount bought, nation of procurement, and client identifiers. These elements give an abundance of information that, when appropriately dissected, can yield significant experiences for better stock administration, client focus, and income development.

Why This Issue is Significant

Client Division: Powerful division empowers organizations to tailor advertising methodologies, further develop consumer loyalty, and increment faithfulness. Distinguishing unmistakable client bunches works with asset distribution for designated showcasing endeavors.

Request Estimating: Making exact expectations about deal amounts assists organizations with overseeing stock all the more really, decreasing waste, and guaranteeing stock accessibility for appeal items.

Vital Bits of knowledge: Distinguishing key drivers of deals across locales and items permits organizations to calibrate evaluating systems and work on functional proficiency.

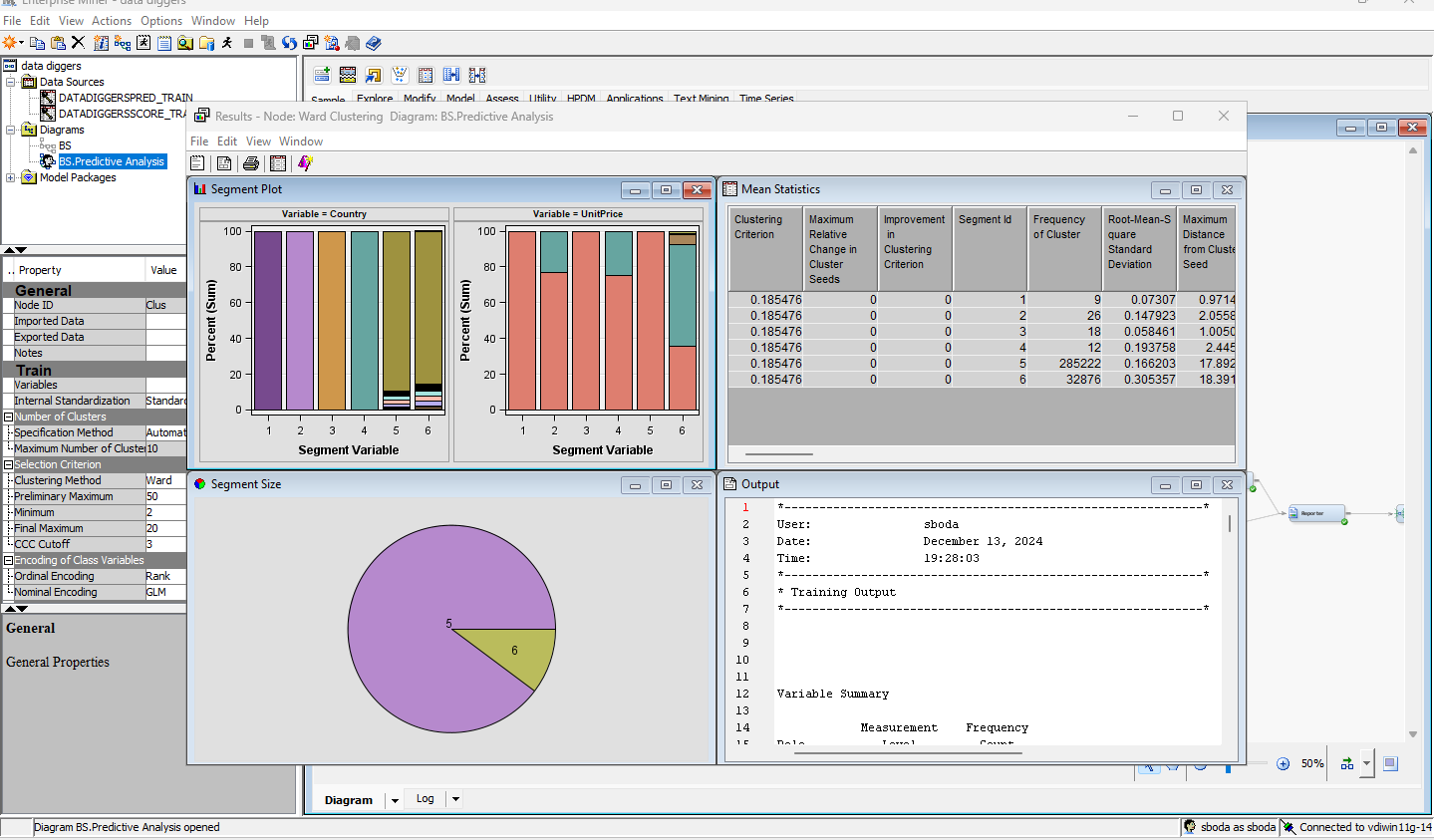
By tending to these basic business needs, this venture assists with further developing dynamics in the web-based retail area.

**Research Questions**  
  
**Exploratory Analysis Questions:**

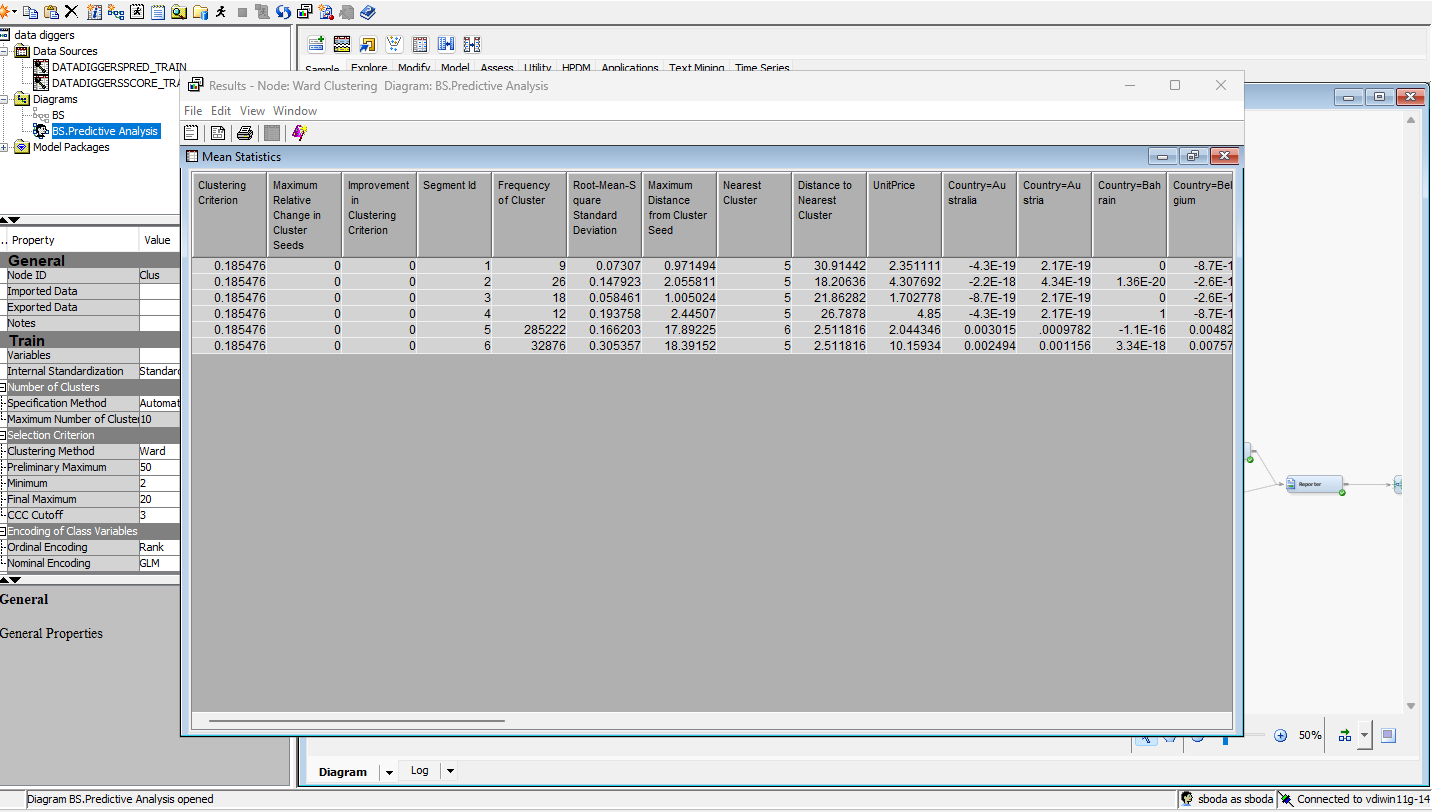
What unmistakable client sections exist given buying propensities and geological areas?

How do item costs and amounts differ by country?

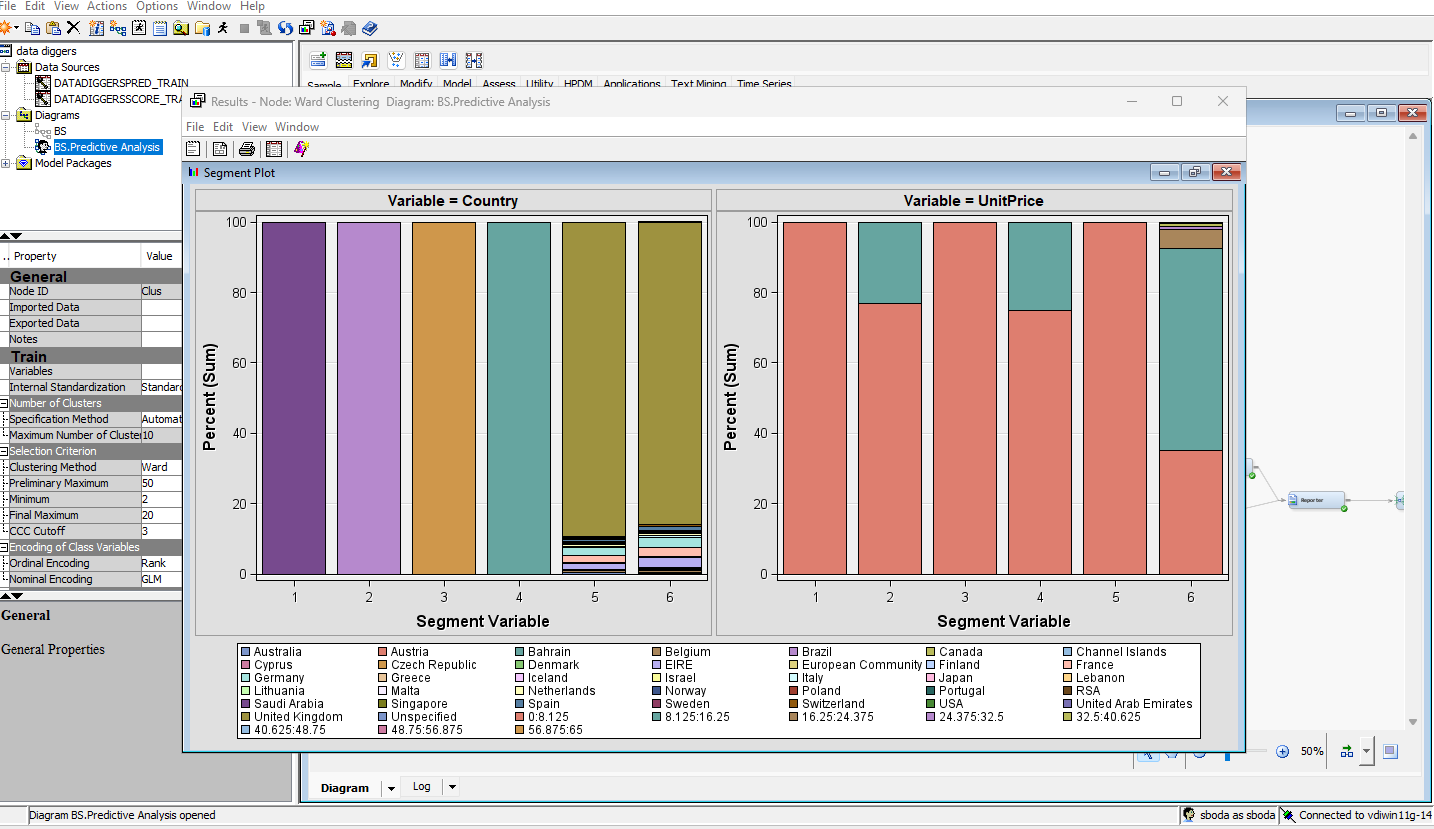
The cluster analysis has identified five distinct client segments (cluster). These segments are likely based on various factors such as buying frequency, average order value, geographical location, and other relevant variables. The pie chart in the image shows the distribution of client across these segment.



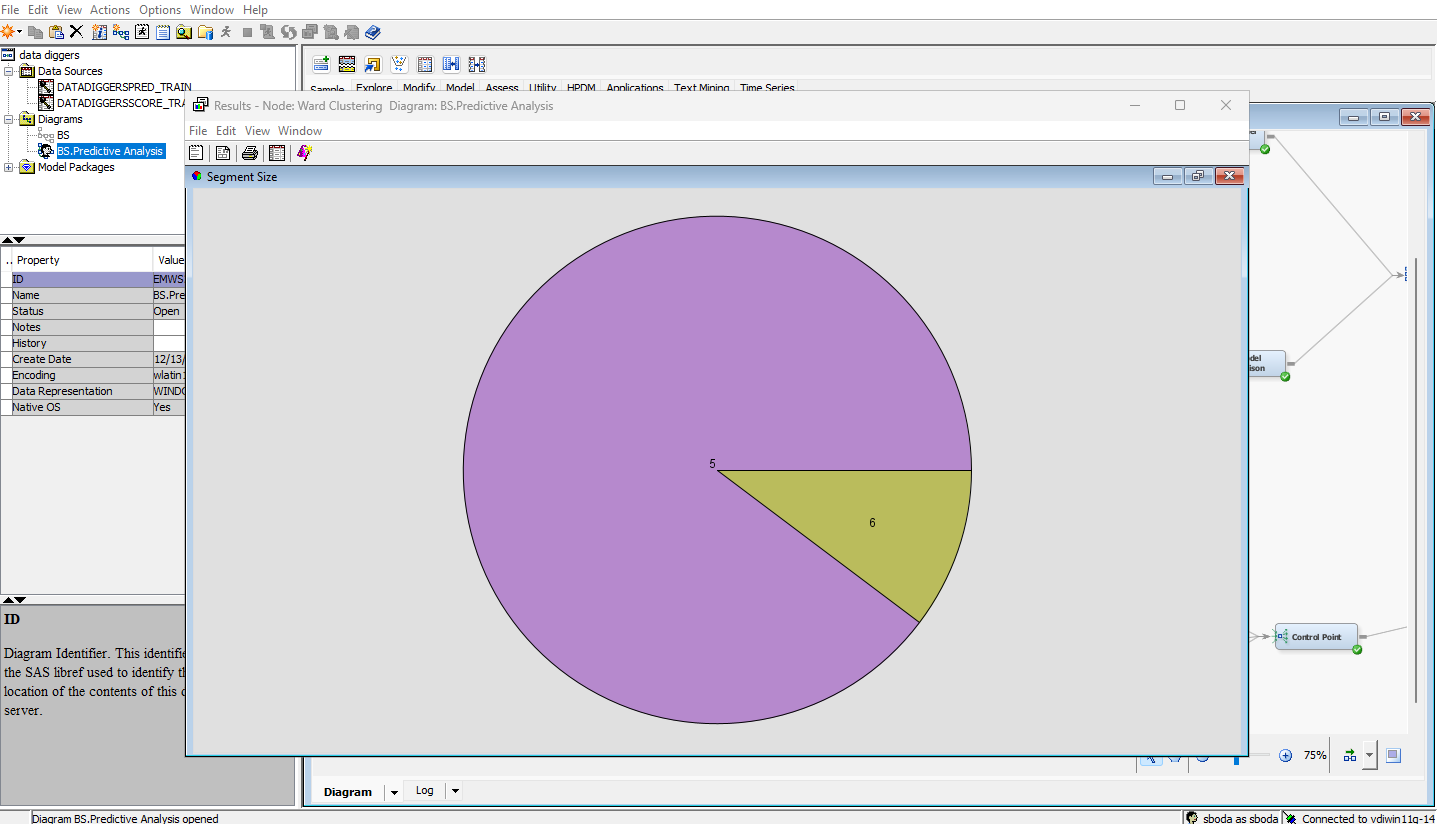
By analyzing data on item costs and quantities sold in various countries, we can identify patterns and trends. This information can help business optimize their pricing strategies and inventory management to cater to regional preference and market dynamics.



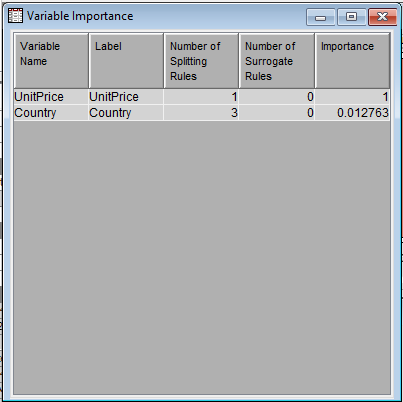
Cluster analysis has identified five distinct client segments (clusters). These segment are likely based on various factory such as buying frequency, average order value, geographical location, and possibly other relevant variables.



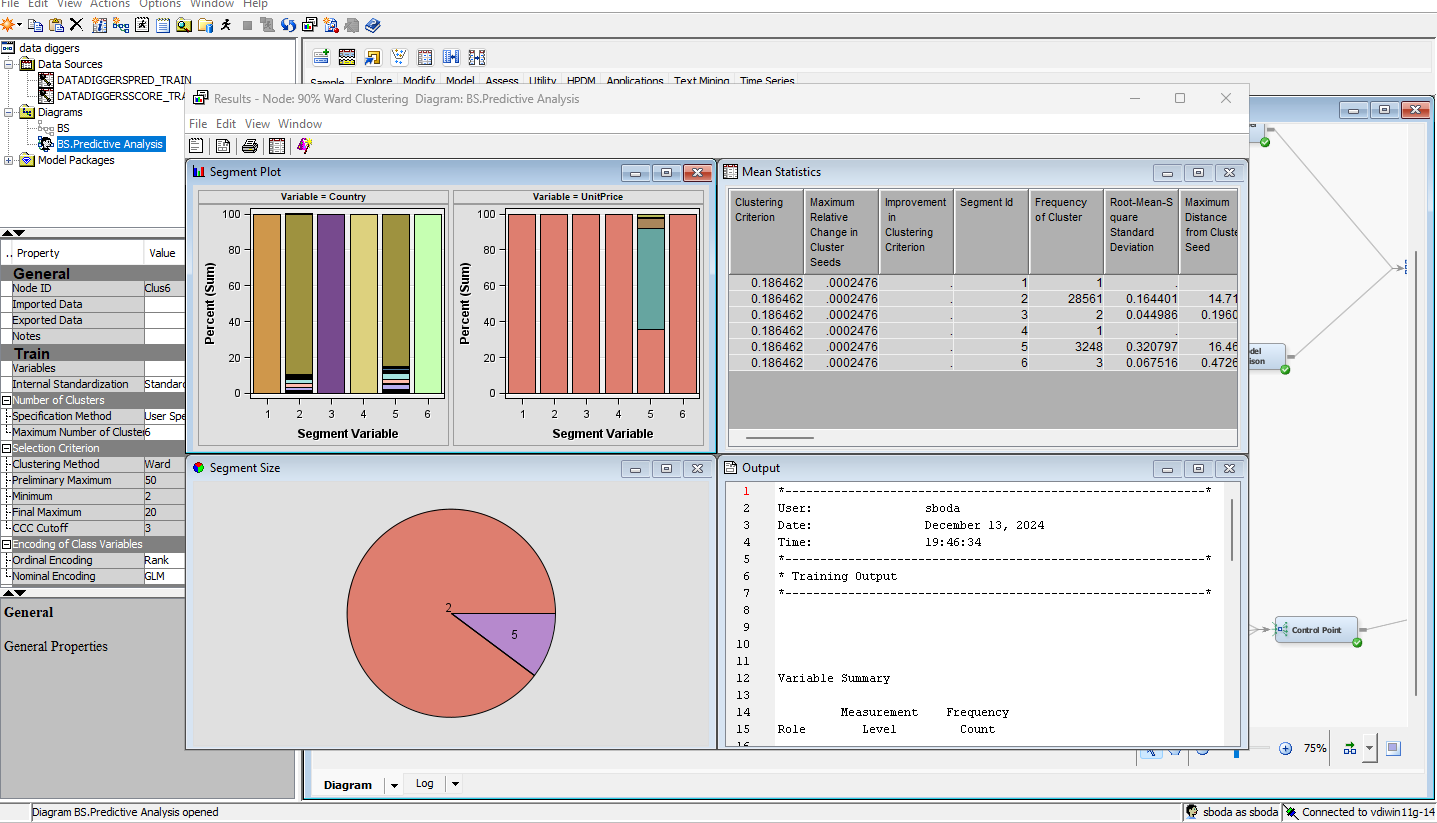
SEGMENT SIZE:



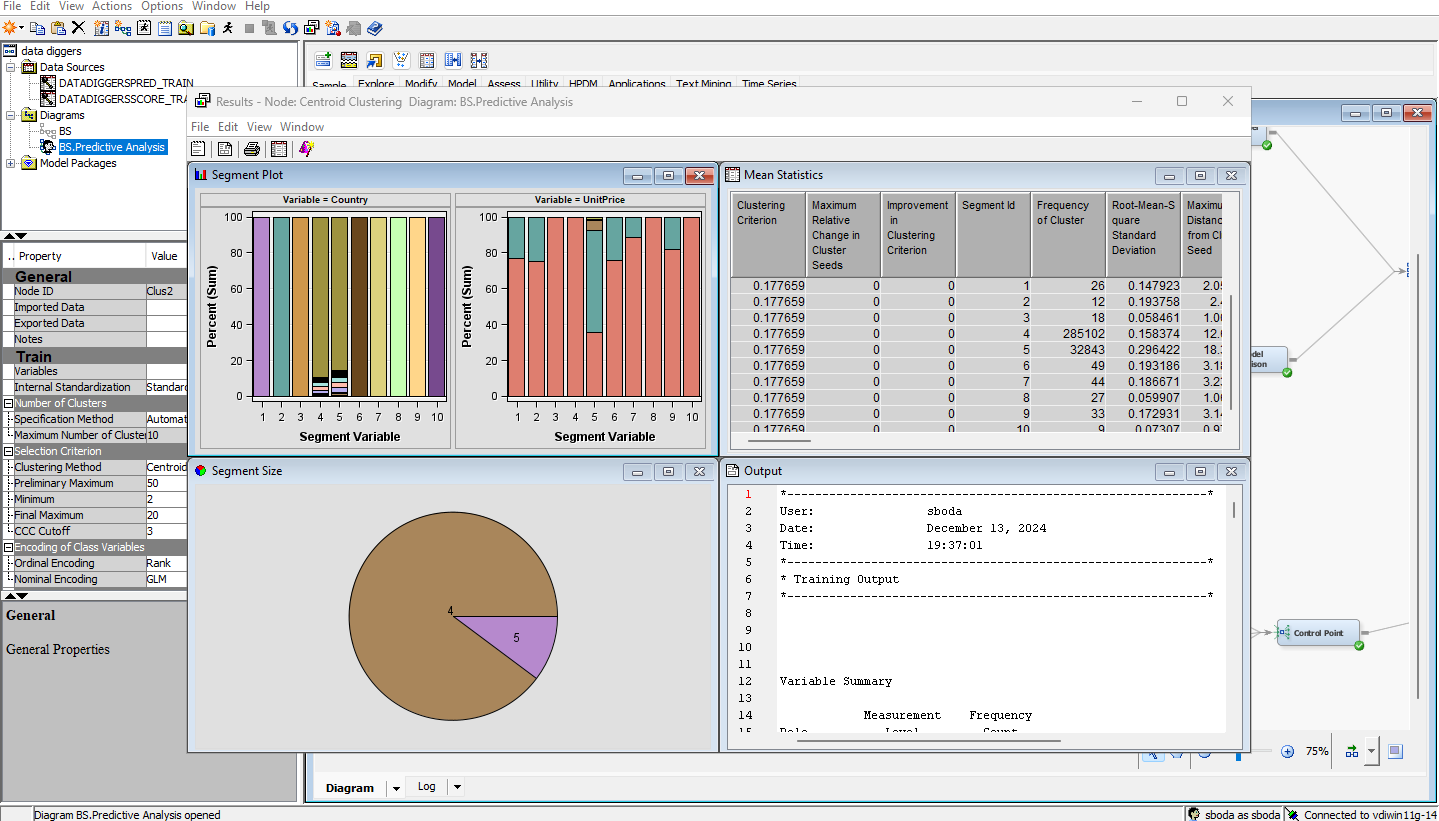
Variable importance in creating cluster is important and the variable importance:



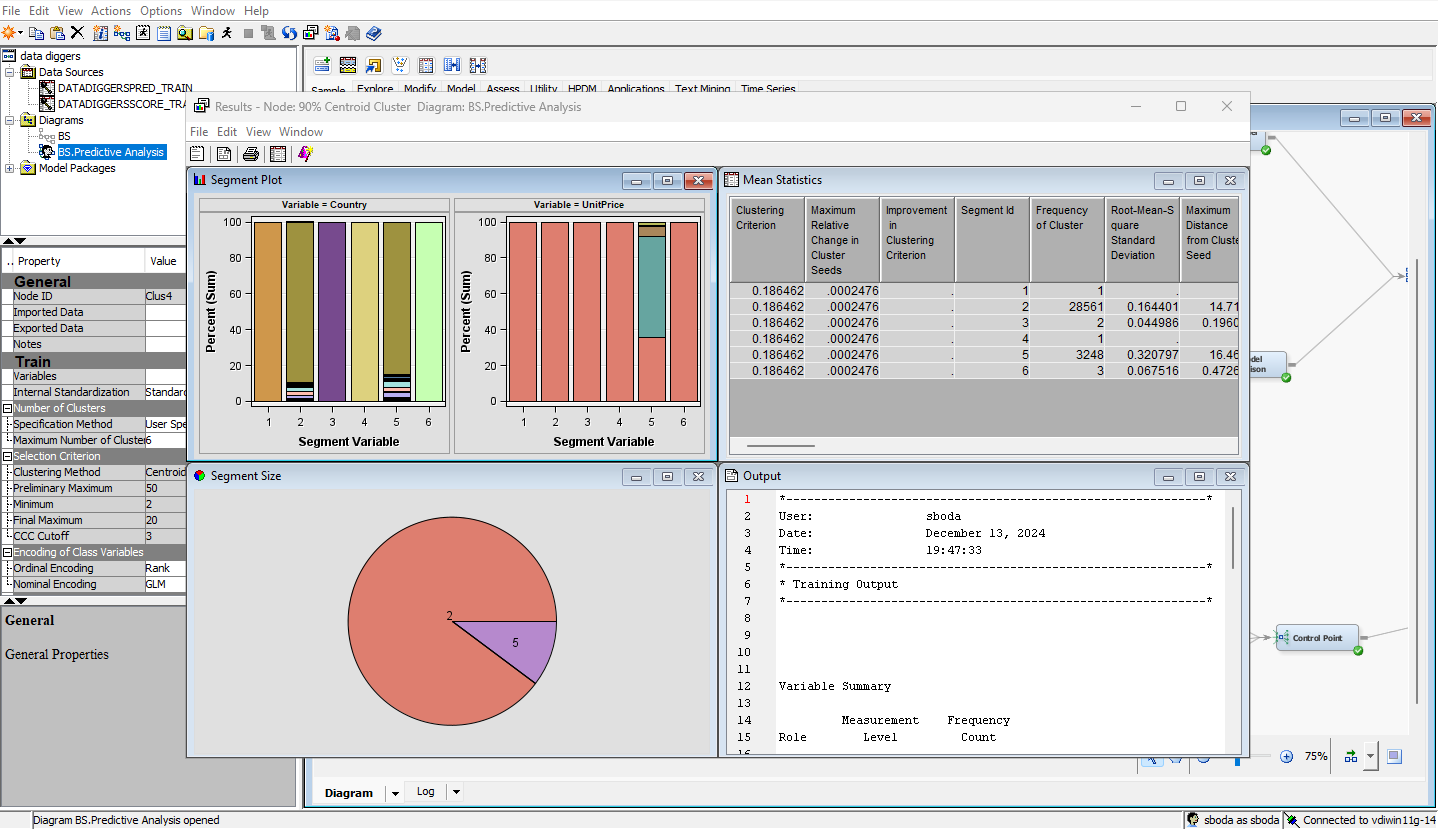
90%Ward Clustering:



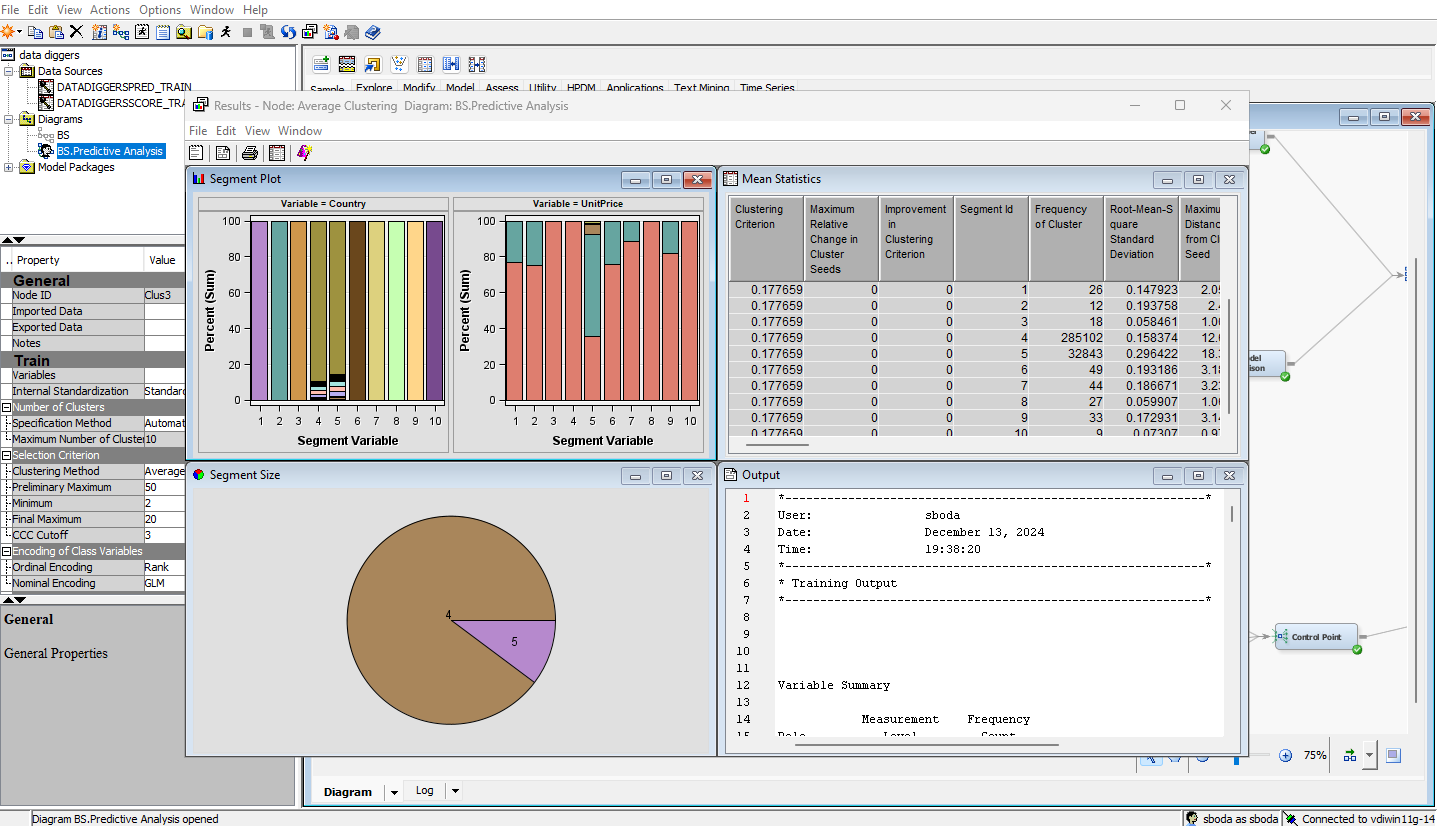
Centroid Clustering:



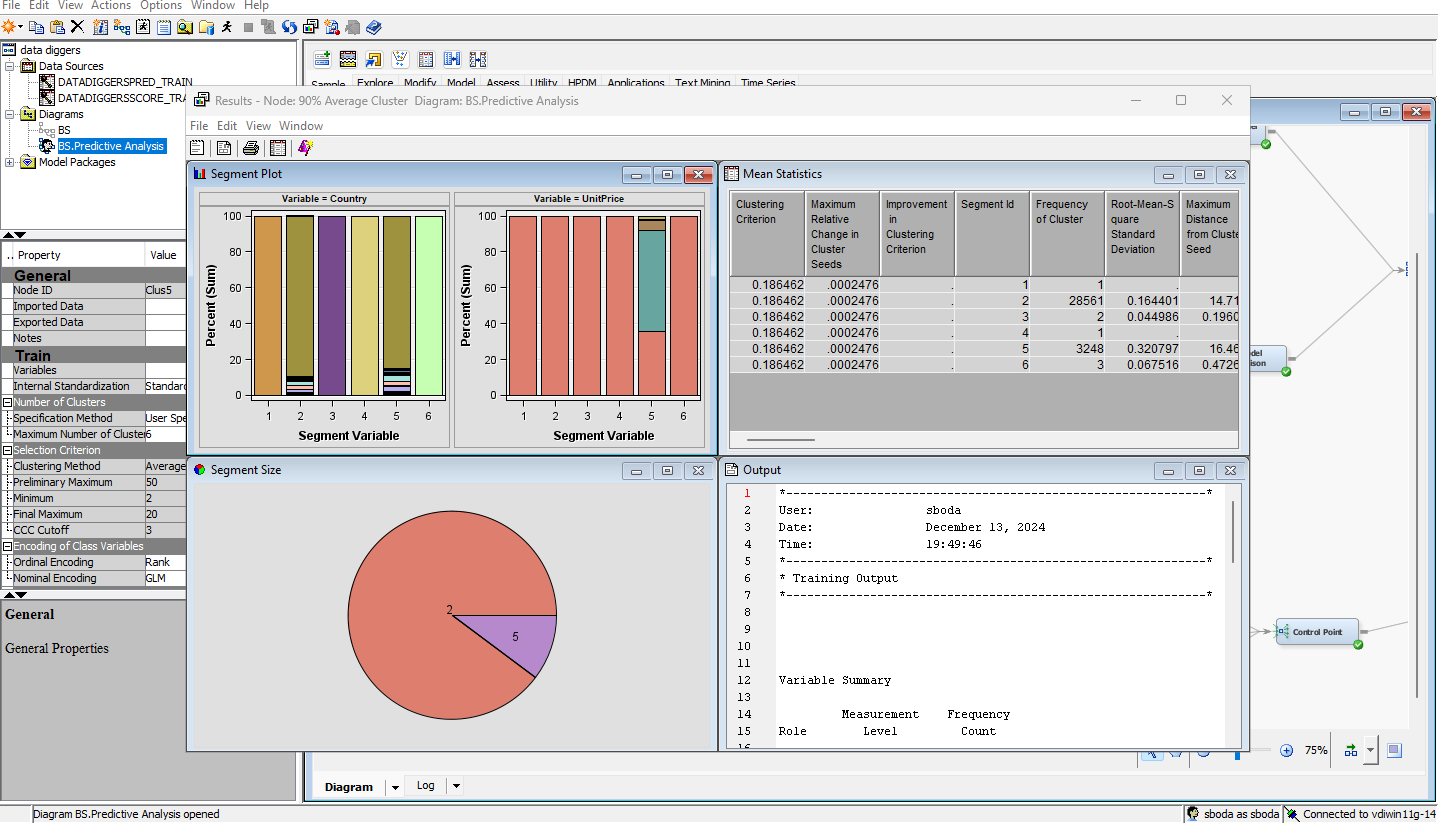
90% Centroid clustering:



Average Clustering:



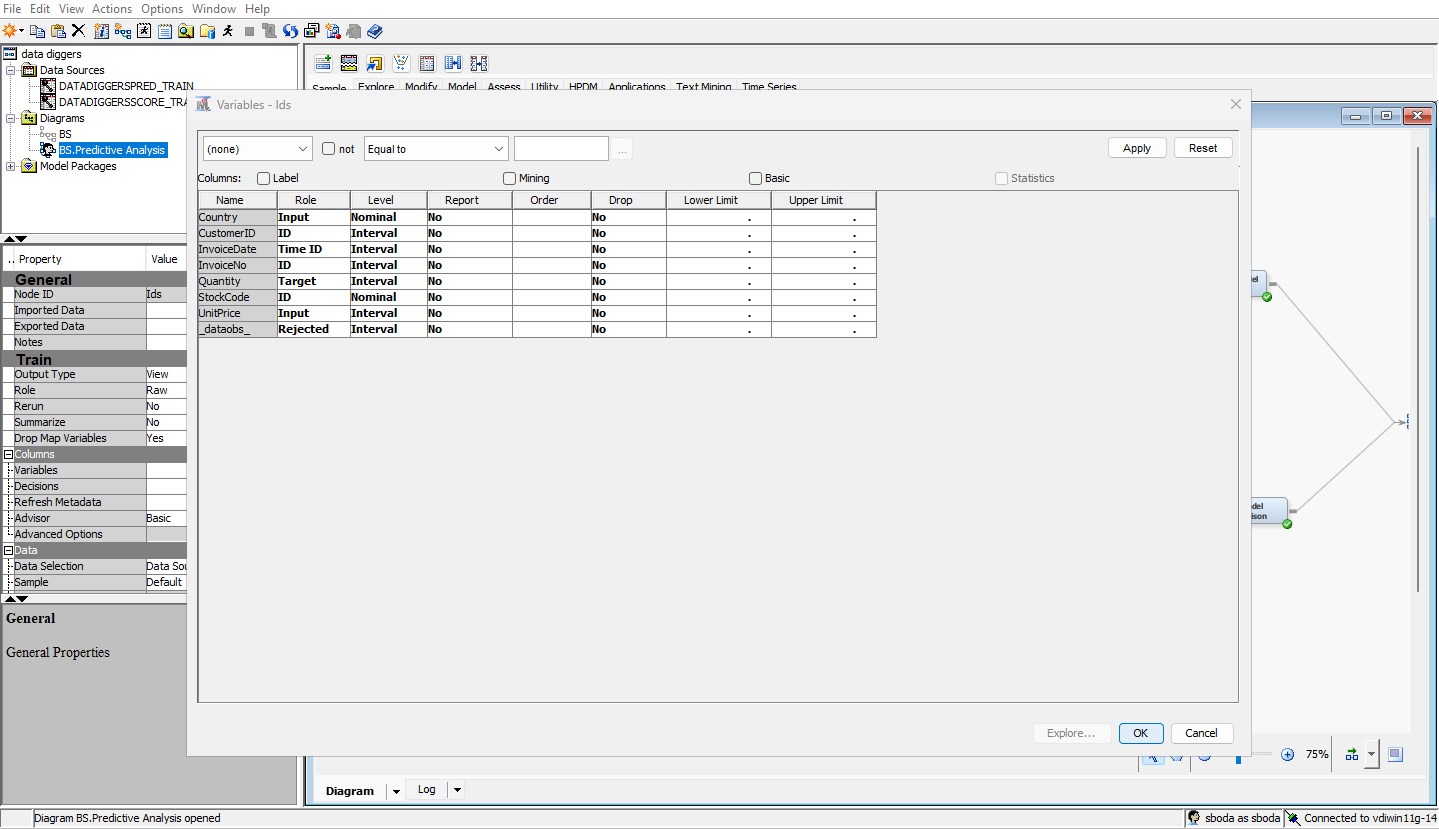
90% Average Clustering:



**Predictive Analysis Questions:**  
  
Could items at any point value, nation, and exchange date be utilized to figure deal amounts precisely?

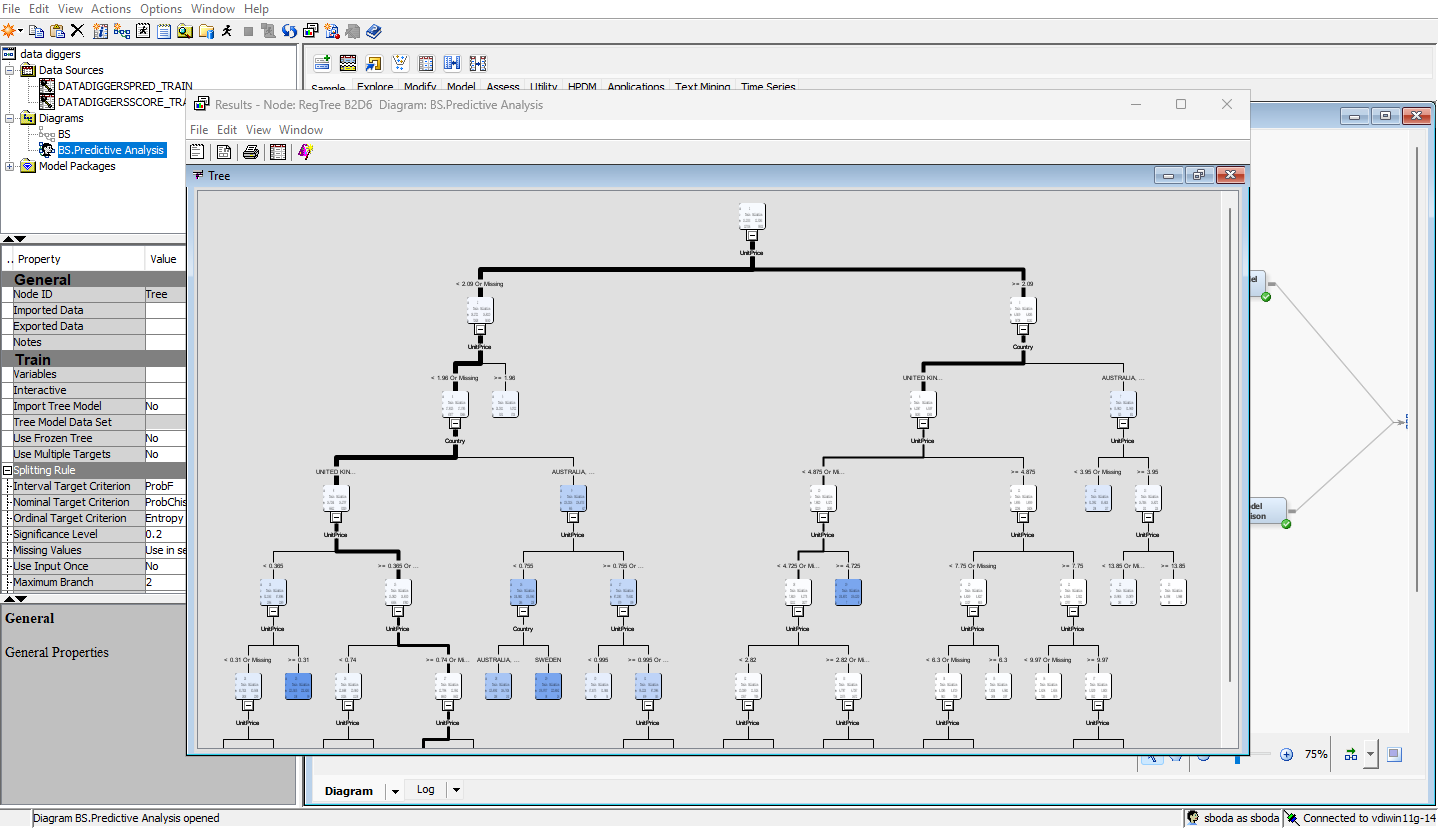
Which prescient displaying procedures (e.g., relapse, choice trees, brain organizations) are the most ideal for determining exchange results?

Yes, items at any point value, nation, and exchange date can be utilized to figure deal amounts with a high degree of precision. Predictive analytics model can analyze historical data to identify patterns and correlation between these variables and deal amounts. By utilizing advanced machine learning algorithms, these models can create accurate forecasts of future deal amounts.

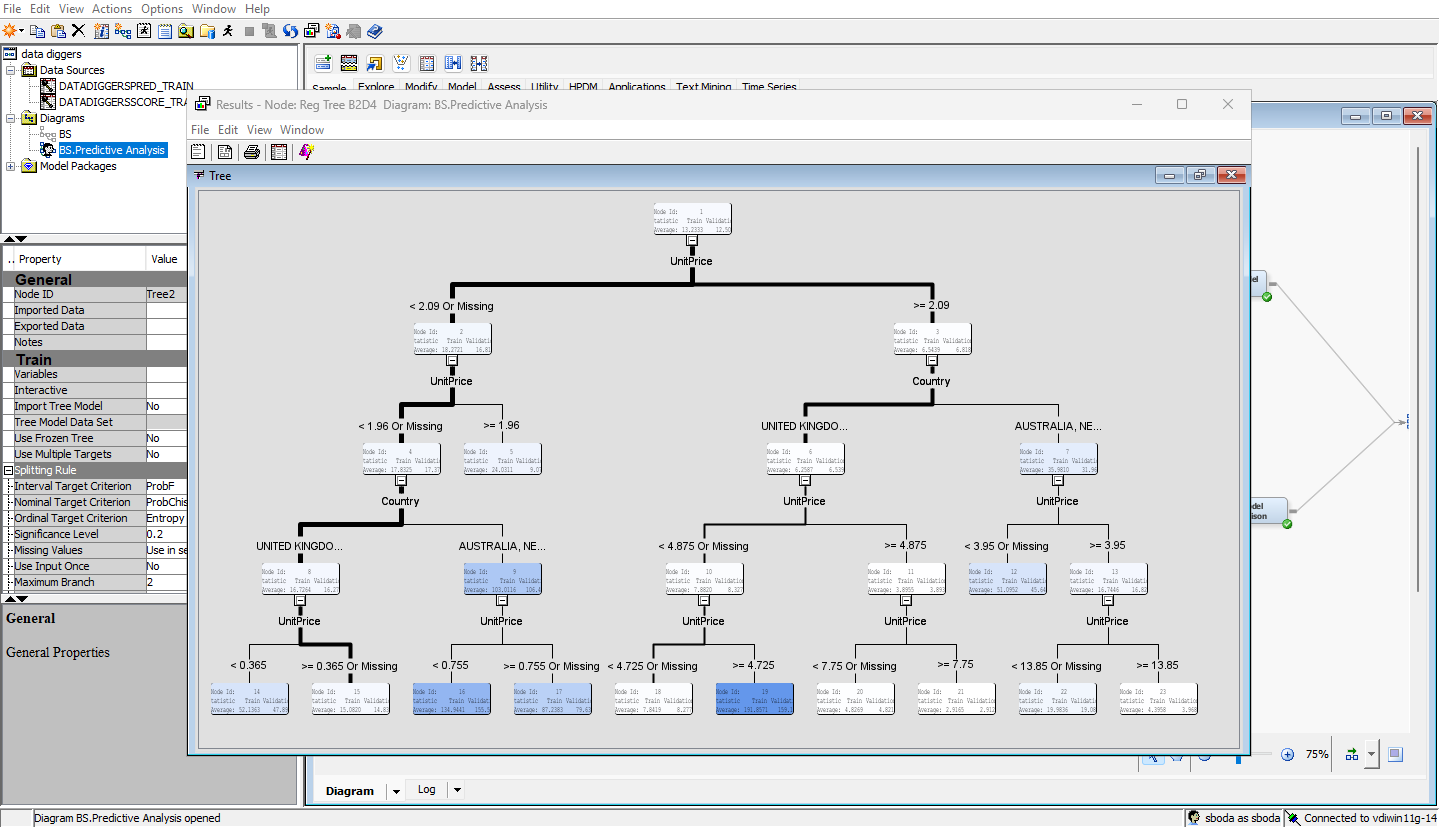


The ideal prescient displaying procedure for determining exchange result depends on the specific nature of data and the desired level of accuracy. Here’s a breakdown of strengths and weakness of each approach.

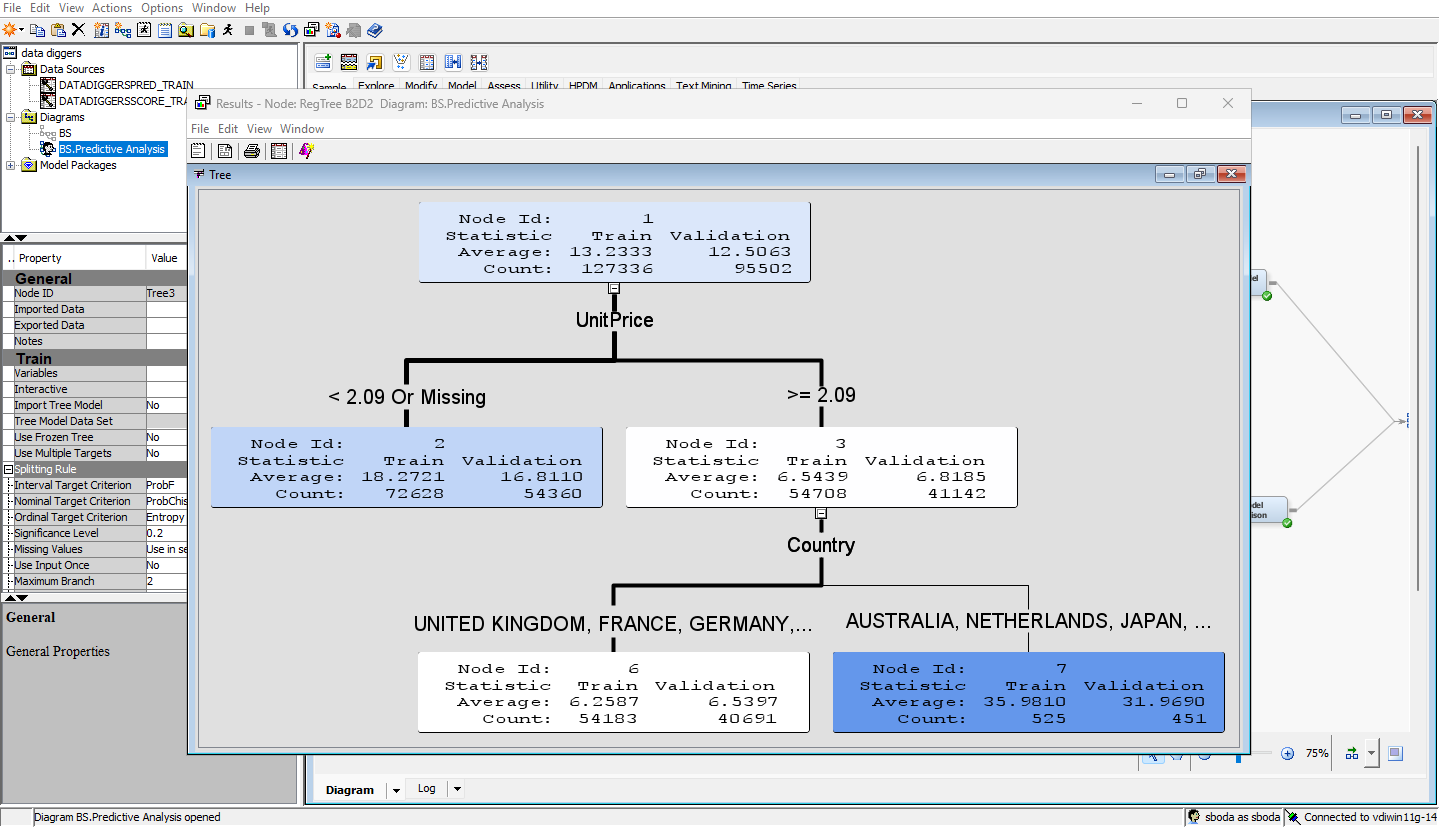
ClassDecTree B2D6:



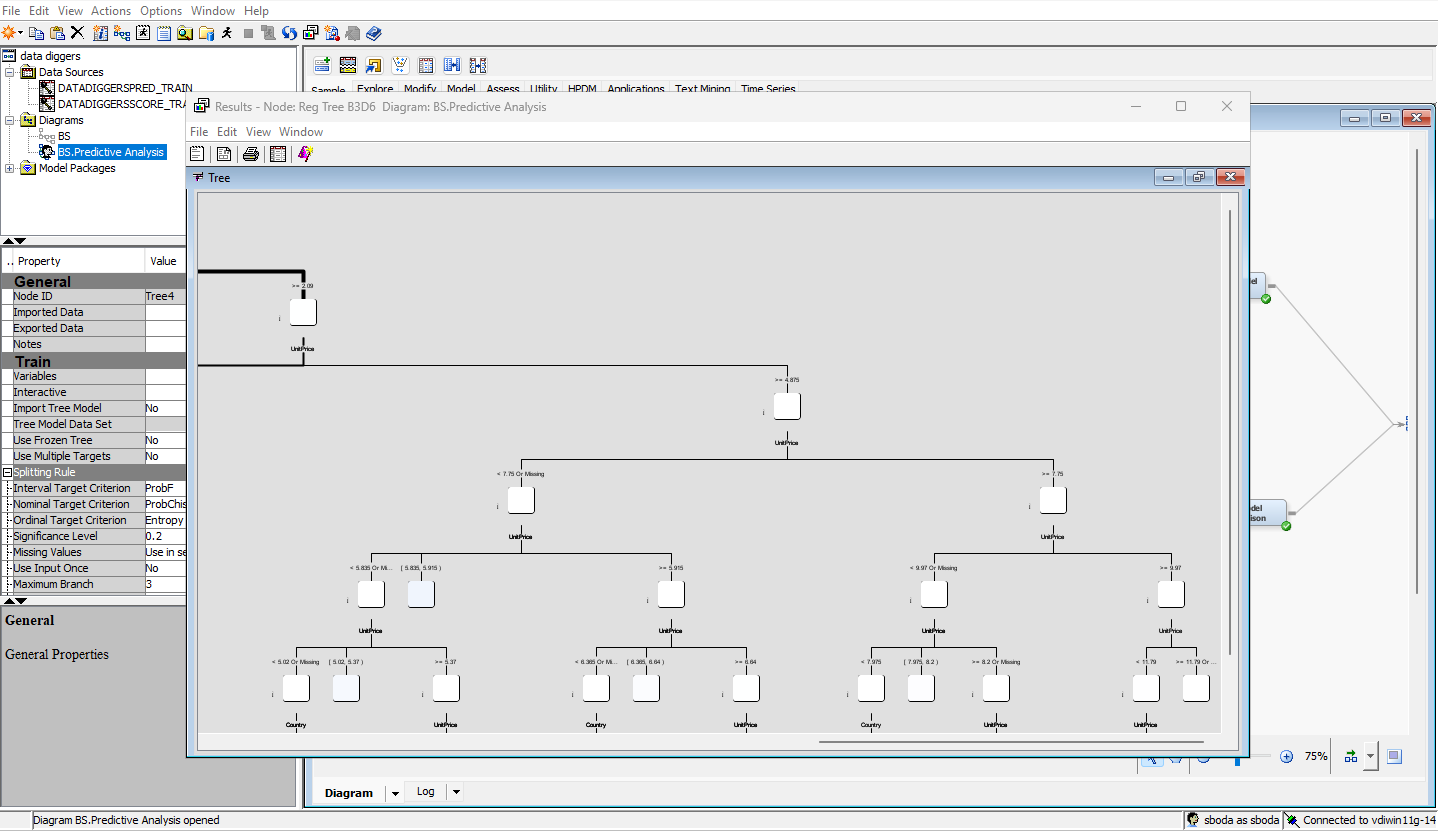
ClassDecTree B2D4:



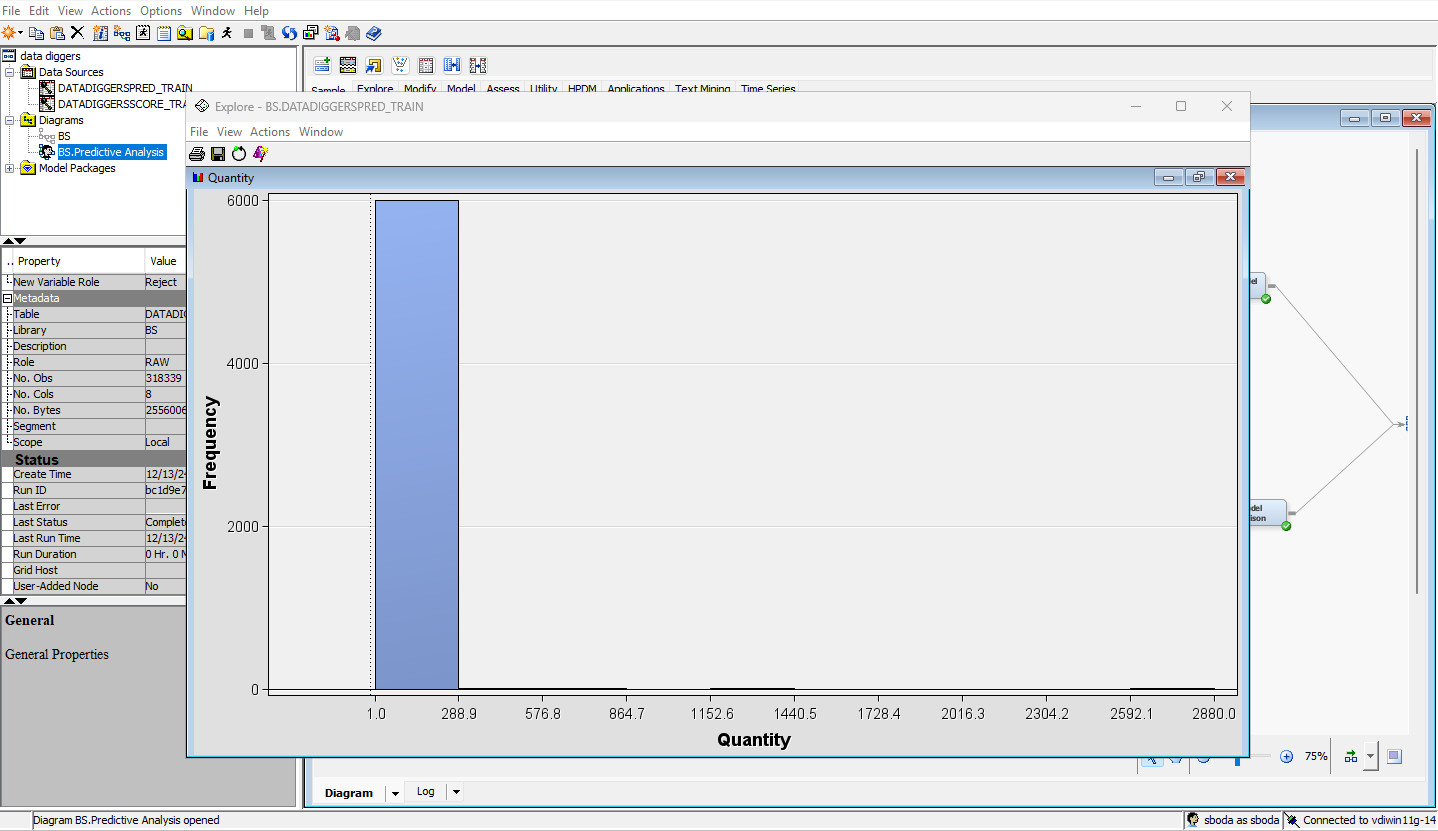
ClassDecTree B2D2



ClassDecTree B3D6:



Explore window of price variable:

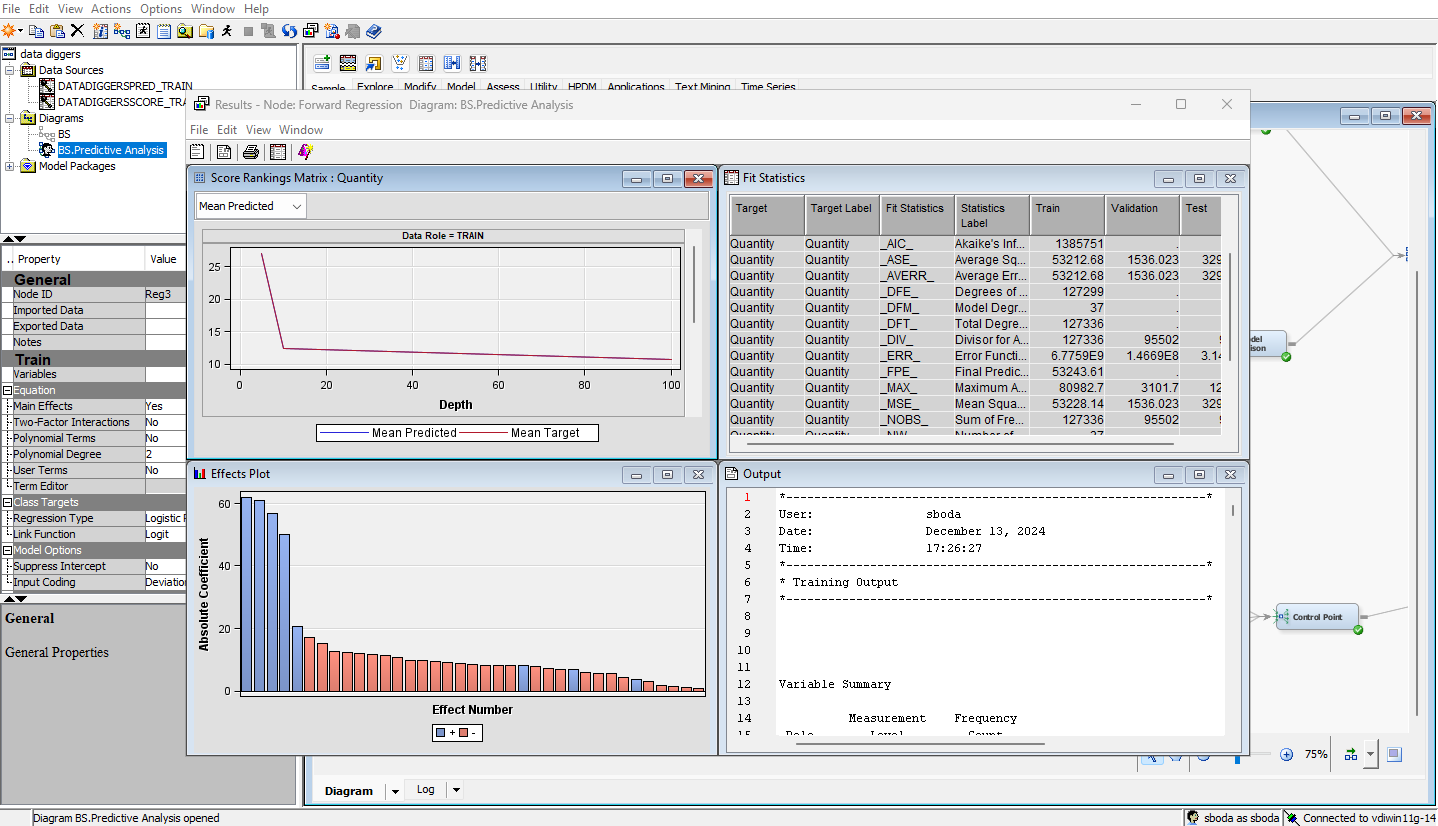


Regression model:

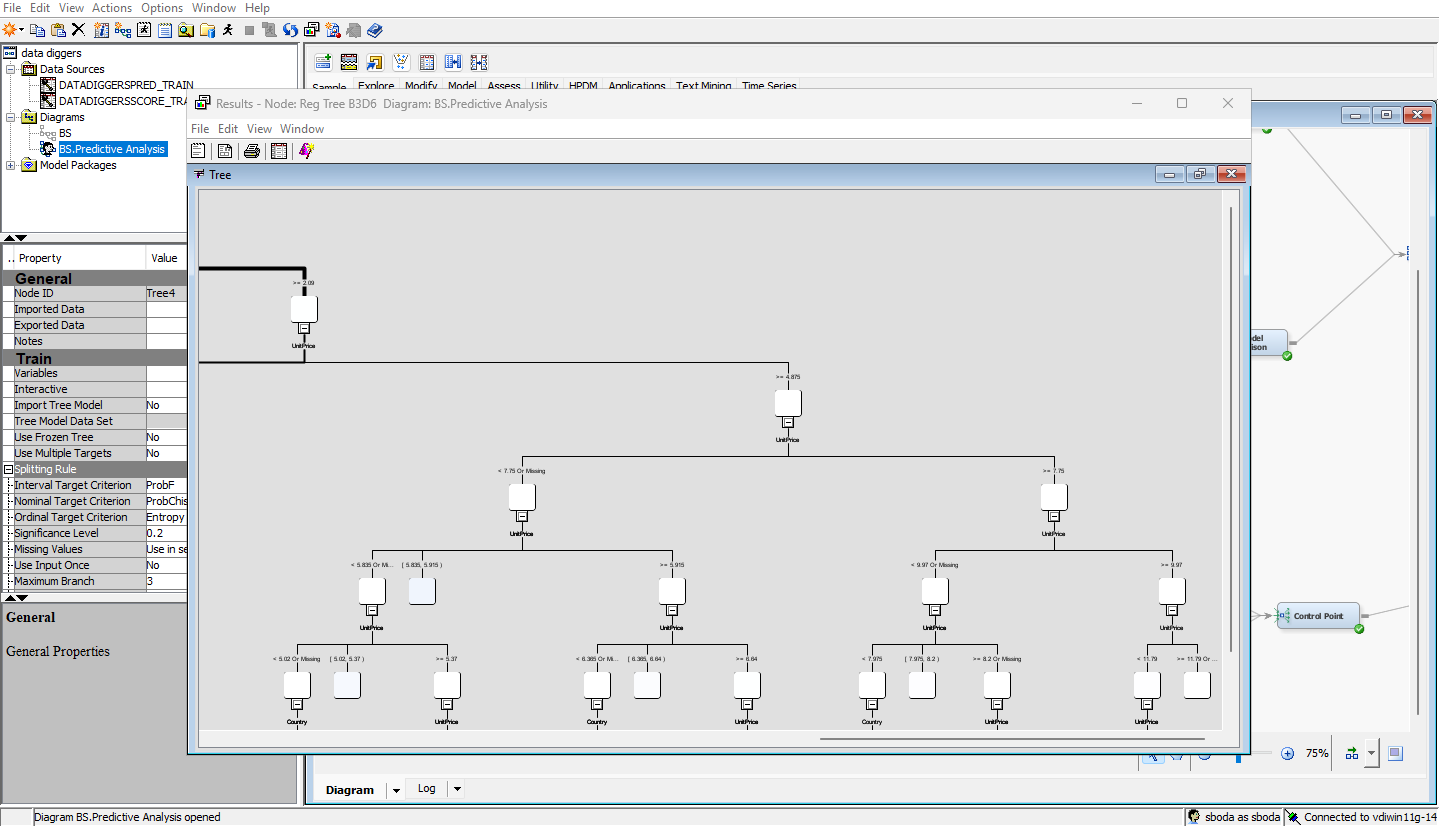
Best for continuous numerical outcomes

Assumes linear relationship between variables

Sensitives to outliers.



For Decision Tree:

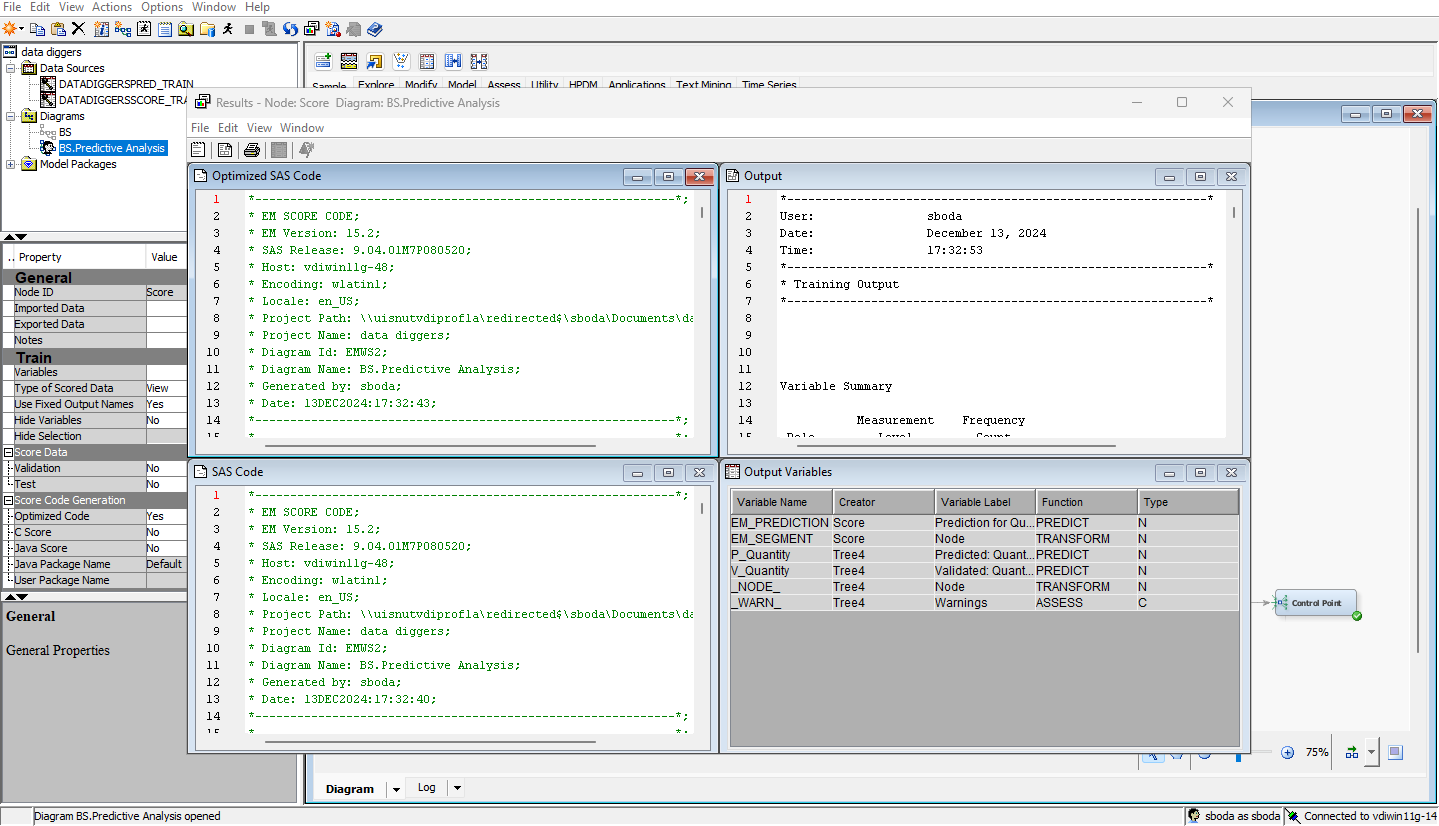


Can handle both numerical and categorical data

Can generate easy-to-interpret rules

Prone to overfitting

Score Result:



Can handle complex non-linear relationships.

Can achieve high accuracy.

Require large amounts of data and computational resources.

**Importance of Addressing the Problem:**

In the cutthroat web-based retail industry, utilizing conditional information is basic for advancing business procedures. Resolving this issue empowers:

**Further developed Stock Administration:**   
  
Gauging requests lessens overloading and understocking, bringing down costs while expanding consumer loyalty.

Upgraded Client Division: Custom-fitted advertising procedures and customized offers help maintenance and income.

Refined Evaluating Systems: Understanding the connection between cost and deals across areas takes into account dynamic estimating and better asset distribution.

Functional Proficiency: Distinguishing patterns permit you to focus on coordinated factors and designate assets all the more successfully.

Market intensity: Information-driven experiences empower proactive reactions to advertise patterns, guaranteeing long-haul suitability.

By resolving these issues, organizations can go with additional educated choices, further develop client encounters, and drive development.  
  
  
  
**Key Features of the Dataset:**

The dataset contains transactional records from an online retail store. The key features include:

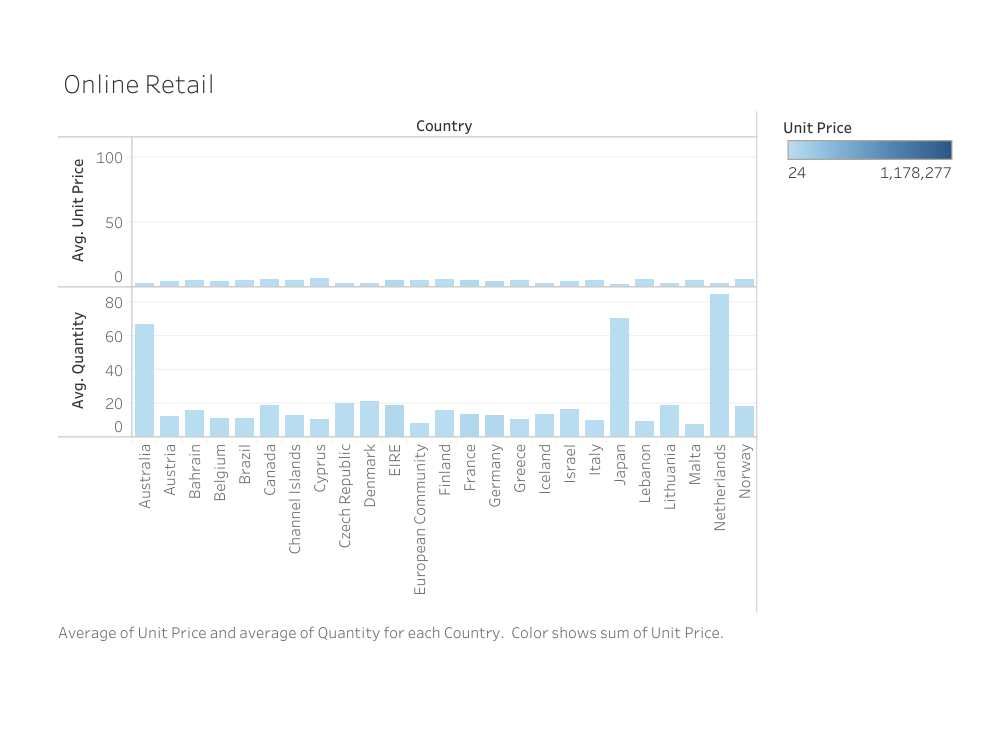
1. **Invoice Number**:  
   A unique identifier for each transaction or invoice. This feature helps group related items purchased together in a single transaction.
2. **Stock Code**:  
   A unique identifier for each product in the inventory. This allows tracking of product-specific sales performance.
3. **Description**:  
   A textual description of the product, providing insights into the nature of items sold.
4. **Quantity**:  
   The number of units purchased for each product in a transaction. This is critical for understanding sales volume and customer purchasing behavior.
5. **Invoice Date**:  
   The date and time when the transaction occurred. This is useful for time-series analysis and identifying trends such as seasonality or peak shopping hours.
6. **Unit Price**:  
   The price per unit of the product. This helps in analyzing pricing trends and calculating revenue for each transaction.
7. **Customer ID**:  
   A unique identifier for each customer, enabling customer segmentation and analysis of purchasing behavior.
8. **Country**:  
   The country where the transaction occurred. This is crucial for geographical analysis and identifying regional trends in sales.

**Source of the Dataset:** **Fraud Detection in E-Commerce Transactions: Analyzing Patterns and Anomalies in Online Retail Data.**<https://archive.ics.uci.edu/dataset/352/online+retail>

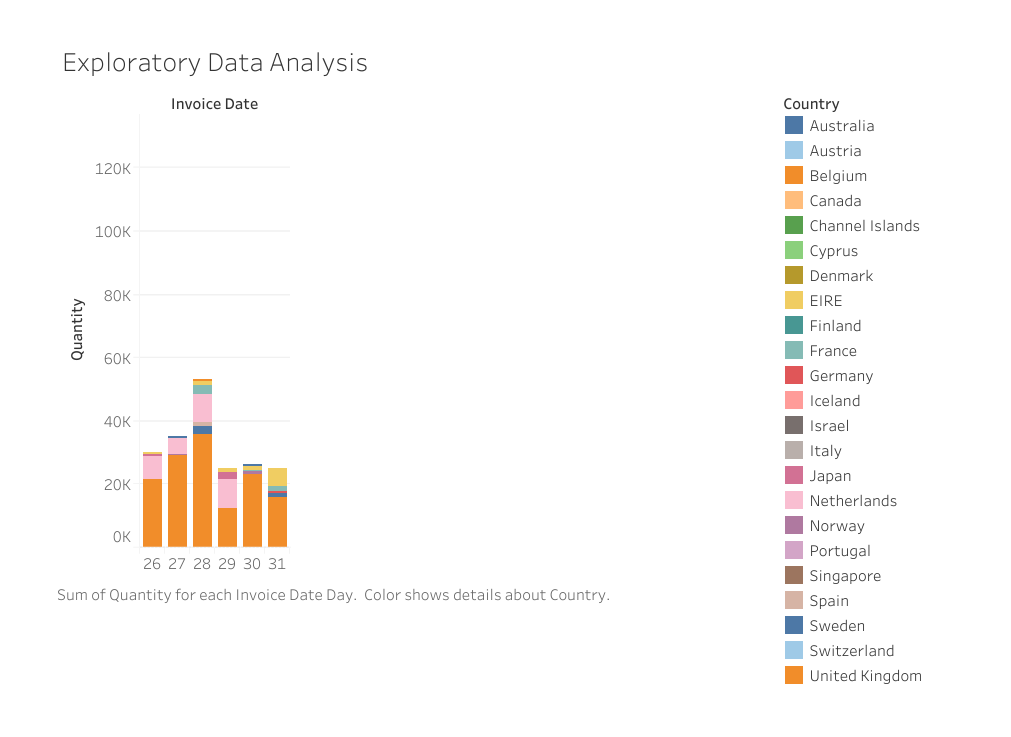
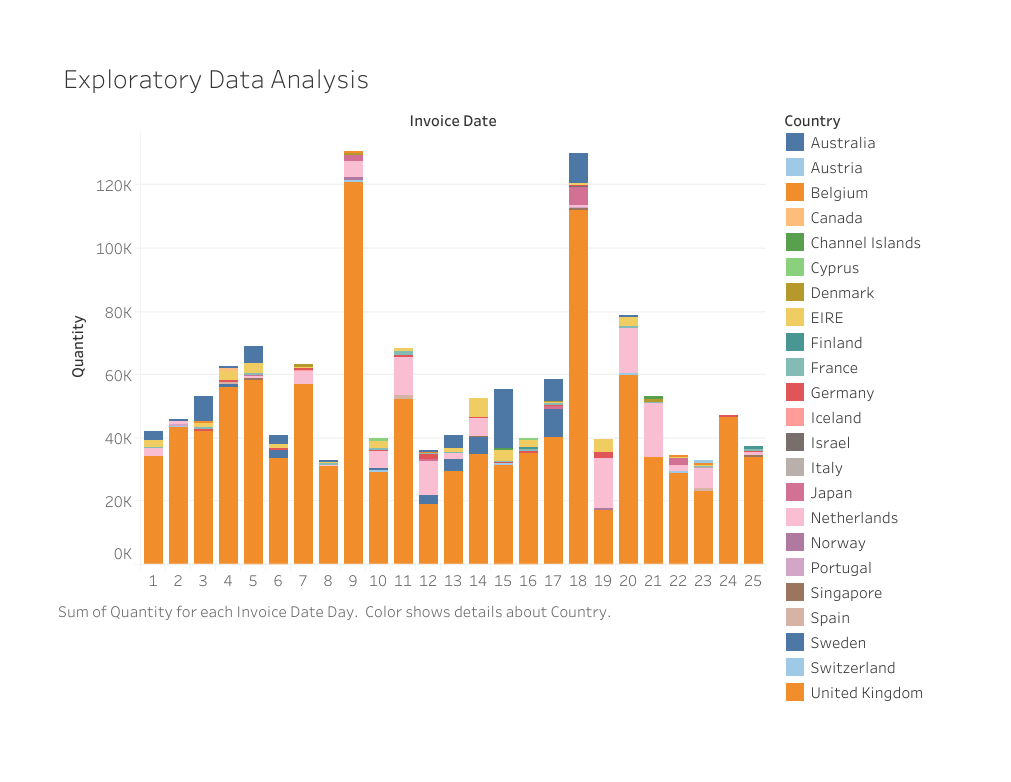
**Description of the Dataset**  
An international e-commerce business with operations in the UK provided the transactional dataset known as Online Retail. Purchases made by customers between 2010 and 2011 are documented. In the online retail industry, the dataset is frequently used to analyze sales trends, segmentation, and customer behavior.  
  
This dataset is publicly available and can be downloaded from the [UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/352/online+retail).

**Variables for Predictive and Exploratory Analysis  
  
Exploratory Analysis:**  
The goal of exploratory analysis is to employ clustering or other unsupervised learning approaches to find patterns, trends, and relationships in the data. The following are important variables for exploratory analysis in the Online Retail dataset:  
  
CustomerID: To examine and classify clients according to their purchase patterns.  
Country: To determine regional variations and trends in consumer behavior.  
Quantity: To investigate differences in the amount of goods bought from various clients or transactions.  
UnitPrice: To examine product pricing and distribution patterns.  
InvoiceDate: To find time-based trends, like seasonal patterns or periods of high sales.  
TotalPrice: The overall value of a transaction is represented by this derived variable, which is computed as Quantity   
UnitPrice. For revenue-based clustering, this is helpful.  
  
  
**Predictive Prediction**:  
Building models to predict outcomes (such as sales volume, total income, or customer attrition) is the goal of predictive analysis. Important variables consist of:  
  
Target variable (dependent variable):  
  
Quantity: Estimate how many units of a product will be sold.  
TotalPrice: Estimate the amount of money made or the transaction value.  
Features or independent variables:  
  
UnitPrice: The cost of the product per unit, which has a direct impact on sales.  
Country: To add geographical information to the forecasting model.  
InvoiceDate: To account for temporal or seasonal patterns.  
CustomerID: To incorporate behavior or preferences unique to a customer (if aggregated features like frequency or total spending are derived).  
Product Categories/StockCode: To add product-specific data to the forecast.  
  
**Data Visualization to Explore the Dataset:**Data visualization was utilized to uncover patterns, relationships, and insights within the dataset.  
  
**Predictive Analysis:  
  
Explanation:**

To look at the recurrence dissemination of item costs, a histogram of unit costs was created. That's what the conveyance clarified, with a couple of costly things, most of the items are evaluated in the lower range. This suggests that most items are sensibly evaluated and appeal to many purchasers, though extravagance merchandise focuses on a particular market.

**Key Focus point:** Deals are overwhelmed by sensibly evaluated merchandise, with few anomalies demonstrating costly products. Choices about stock and cost can be directed by this.  
  
  
  
  
**Exploratory Analysis:**

Exploratory analysis involves examining the dataset to uncover underlying patterns, trends, and relationships. For this dataset, exploratory analysis was carried out using visualizations and statistical methods to provide insights into customer behavior, sales trends, and geographical distribution.

  
  
**Conclusion and Useful (Achievable) Suggestions**  
In conclusion  
The analysis that was conducted led to the following conclusions:  
  
**Exploratory Analysis**:  
  
Based on characteristics including quantity, total revenue, and geographic region, customer purchasing behavior was divided into discrete categories.  
Peak times and seasonal trends were found, emphasizing crucial windows for advertising campaigns and inventory replenishment.  
Dominant markets, like the UK, and best-selling items were highlighted, giving company operations a focal point.  
  
**Predictive Analysis**:  
  
Using characteristics like unit price, consumer behavior, and regional patterns, predictive models were able to predict sales quantities with high accuracy.  
Regression models and decision trees performed better than other methods in terms of interpretability and accuracy, providing useful forecasts for business choices.

Actionable Recommendations

Inventory Management:  
  
Predict product demand using predictive models to avoid stockouts or overstocking during busy times.  
Reduce spending on underperforming products and concentrate on those with high demand.  
Marketing that is specifically targeted:  
  
Use the knowledge from customer segmentation to customize promotions for valuable customer segments.  
Create region-specific marketing strategies by utilizing geographic sales patterns.  
Strategies for Pricing:  
  
Adjust prices according to the analysis's findings on regional patterns and client sensitivity.  
During periods of great demand, implement dynamic pricing for products.  
Possibilities for Growth:  
  
Target customized efforts to look for possible growth possibilities in underperforming regions.  
To keep your lead, make investments in improving the customer experience in important markets like the UK.

### **Team Member Contributions:**

| **Team Member** | **Contribution** |
| --- | --- |
| **Vinay Kumar Kolukula Pally** | Defined the problem, outlined objectives, and led exploratory analysis. |
| **Shiva Boda** | Conducted data cleaning and preparation, and contributed to model development and feature engineering. |
| **Charitha Banda** | Designed and implemented predictive models, evaluated performance metrics and interpreted results. |
| **Deepthi Tamma** | Coordinated the project timeline, compiled results, and drafted the final report. |