Wholesale customer segmentation

Problem Description :

Background: The wholesale industry deals with the distribution of goods and products from manufacturers to retailers, restaurants, hotels, and other businesses that require large quantities of goods. Understanding the purchasing behavior of wholesale customers is crucial for wholesalers to optimize their supply chain, inventory management, and marketing strategies.

Dataset Information: The dataset used for wholesale customer segmentation contains information on various products purchased by different customers in the wholesale industry. Each customer's purchasing behavior is recorded in terms of the amounts spent on different product categories. The dataset also includes information about the region and channel (Horeca or Retail) to which the customers belong.

Features:

1. Fresh: Annual spending (in monetary units) on fresh products.
2. Milk: Annual spending (in monetary units) on milk products.
3. Grocery: Annual spending (in monetary units) on grocery products.
4. Frozen: Annual spending (in monetary units) on frozen products.
5. Detergents_Paper: Annual spending (in monetary units) on detergents and paper products.
6. Delicatessen: Annual spending (in monetary units) on delicatessen products.
7. Region: Region to which the customer belongs (Oporto, Lisbon, or Other).
8. Channel: Type of channel through which the products are purchased (Horeca or Retail).

Problem Statement: The main objective of this project is to perform wholesale customer segmentation using clustering techniques. The goal is to group similar customers based on their purchasing behavior, which will help wholesalers in tailoring their strategies and services to meet the specific needs of different customer segments.

Approach: The project involves the following steps:

1. Data Preprocessing: Load the dataset, handle missing values (if any), and convert categorical variables into meaningful data.
2. Data Transformation: As the data is not normally distributed, apply Box-Cox transformation to make it more normally distributed. Scale the data using Min-Max scaling to bring all features to a similar scale.
3. Handling Outliers: Remove outliers using the Interquartile Range (IQR) method to ensure the clustering is not affected by extreme values.
4. Checking for Normality: Verify the normality of the transformed data to ensure it meets the assumptions of clustering algorithms.
5. Clustering with K-Means: Use the K-Means clustering algorithm to group customers into segments based on their purchasing behavior. Determine the optimal number of clusters using the Elbow method.
6. Principal Component Analysis (PCA): Perform PCA to reduce the dimensionality of the data and visualize the clusters in a 2D space.
7. Clustering with Hierarchical Clustering: Apply Hierarchical Clustering to further validate the customer segments.
8. Cluster Analysis: Analyze the characteristics of each cluster, such as average spending on different product categories, and interpret the customer segments.

Outcome: By performing wholesale customer segmentation, wholesalers can gain valuable insights into customer preferences and needs. This will enable them to develop targeted marketing strategies, optimize product offerings, and improve overall customer satisfaction. The identified customer segments can be used for personalized marketing campaigns and to enhance the efficiency of inventory management and supply chain operations.

Possible Framework :

1. **Introduction:**
- Briefly explain the background of the project and the importance of wholesale customer segmentation.
- Provide an overview of the dataset and its features.
2. **Data Preprocessing:**
- Load the dataset and check for missing values. If present, handle missing values appropriately.
- Convert categorical variables (Region and Channel) into meaningful data using mapping.
- Explore the statistical summary of the dataset to gain insights.
3. **Data Transformation:**
- Check the distribution of the numeric variables in the dataset.
- Apply the Box-Cox transformation to make the data more normally distributed.
- Use Min-Max scaling to bring all features to a similar scale for clustering.
4. **Handling Outliers:**
- Detect and remove outliers using the Interquartile Range (IQR) method to ensure accurate clustering.
5. **Checking for Normality:**
- Verify the normality of the transformed data to ensure it meets the assumptions of clustering algorithms.
- Optionally, use quantile-quantile plots to visualize the normality of the features.
6. **Clustering with K-Means:**
- Implement the K-Means clustering algorithm to group customers into segments based on their purchasing behavior.
- Determine the optimal number of clusters using the Elbow method.
- Fit the K-Means model to the data and obtain cluster assignments.
7. **Principal Component Analysis (PCA):**
- Perform PCA to reduce the dimensionality of the data and visualize the clusters in a 2D space.
- Plot the explained variance ratio to determine the number of principal components to use.
8. **Clustering with Hierarchical Clustering:**
- Apply Hierarchical Clustering to further validate the customer segments.

- Plot the dendrogram to visualize the hierarchical clustering process.

9. Cluster Analysis:

- Analyze the characteristics of each cluster, such as average spending on different product categories.
- Interpret the customer segments to understand their distinct preferences and needs.

10. Visualizations:

- Create visualizations to represent the clusters and their characteristics, such as bar plots and scatter plots.

11. Conclusion:

- Summarize the findings from the wholesale customer segmentation process.
- Discuss the insights gained from the clusters and how they can be used to enhance business strategies.

12. Future Recommendations:

- Suggest possible future work and improvements for the project.
- Mention how the customer segments can be leveraged to optimize inventory management, marketing, and supply chain operations.

13. References:

- Provide references to any external sources or libraries used in the project.

Code Explanation :

*If this section is empty, the explanation is provided in the .ipynb file itself.

1. Importing Libraries: The code begins by importing several Python libraries that will be used throughout the project. These libraries include numpy, pandas, plotnine, matplotlib, seaborn, sklearn, scipy, and statsmodels. Each library serves specific purposes, such as data manipulation, visualization, statistical analysis, and clustering algorithms.

2. Data Preprocessing: The next step is to load the dataset into a pandas DataFrame. The dataset contains information about wholesale customers and their spending habits on different product categories. The code then checks for missing values in the dataset and handles them if present. Additionally, the categorical variables "Region" and "Channel" are mapped to meaningful labels for better analysis.

3. Data Transformation: Before performing clustering, it is crucial to ensure that the data is normally distributed and scaled appropriately. To achieve this, the code applies the Box-Cox transformation to make the numeric variables more normally distributed. Next, the Min-Max scaling technique is used to bring all the features to a similar scale, which is necessary for accurate clustering.

4. Handling Outliers: Outliers can significantly affect the clustering process and result in inaccurate segmentation. Therefore, the code detects and removes outliers using the Interquartile Range (IQR) method, ensuring the data is robust and reliable for clustering.

5. Checking for Normality: To validate the success of the data transformation, the code checks for the normality of the transformed data. It is essential for clustering algorithms that the data meets the assumptions of normality. The normality is verified using statistical tests and quantile-quantile plots.

6. Clustering with K-Means: The primary clustering algorithm used in the project is K-Means. The code implements the K-Means algorithm and determines the optimal number of clusters using the Elbow method. The K-Means model is then fitted to the data, and each data point is assigned to a specific cluster based on their purchasing behavior.

7. Principal Component Analysis (PCA): PCA is a dimensionality reduction technique used to visualize the clusters in a two-dimensional space. The code performs PCA to reduce the dimensionality of the data and obtain the principal components that explain most of the variance in the dataset.

8. Clustering with Hierarchical Clustering: In addition to K-Means, the code also applies Hierarchical Clustering to validate the customer segments obtained from K-Means. Hierarchical Clustering is a different approach to grouping data points based on their similarity. The code plots a dendrogram to visualize the hierarchical clustering process.

9. Cluster Analysis: After clustering, the code analyzes the characteristics of each cluster. For example, it calculates the average spending of each cluster on different product categories. This analysis helps to understand the distinct preferences and needs of each customer segment.

10. Visualizations: The code creates various visualizations to represent the clusters and their characteristics. These visualizations include bar plots and scatter plots, which aid in interpreting and communicating the results effectively.

11. Conclusion: The code concludes by summarizing the findings from the wholesale customer segmentation process. It highlights the insights gained from the clusters and how they can be utilized to enhance business strategies, such as inventory management, marketing, and supply chain operations.

12. Future Recommendations: The code includes future work and recommendations for the project, suggesting potential improvements and applications of the customer segmentation results.

13. References: Lastly, the code provides references to any external sources or libraries used in the project.

Overall, the code follows a systematic workflow, starting from data preprocessing and transformation to clustering and visualization of the results. It aims to provide meaningful insights into customer behavior and preferences, helping businesses make data-driven decisions to optimize their operations and better serve their customers.

Future Work :

Introduction: The future work for the Wholesale Customer Segmentation project aims to improve the existing clustering analysis and explore additional insights from the dataset. It involves refining the data preprocessing, experimenting with different clustering algorithms, and conducting further analysis to derive more valuable business insights.

Step-by-Step Guide for Future Work:

1. Enhanced Data Preprocessing:

- **Feature Engineering:** Consider exploring new features or transforming existing ones to better capture customer behavior and spending patterns. For example, create aggregated features like total spending, spending ratio, or spending trends over time.
- **Handling Skewed Data:** Explore alternative data transformation techniques for normalization, such as log transformation or power transformation, to deal with skewed data distributions more effectively.
- **Handling Outliers:** Experiment with different outlier detection methods, like Z-score or Mahalanobis distance, to better handle outliers in the dataset.

2. Experiment with Different Clustering Algorithms:

- **K-Means Variants:** Explore other variants of the K-Means algorithm, such as Mini-Batch K-Means or K-Medoids, to compare their performance and identify the most suitable algorithm for the dataset.
- **Density-Based Clustering:** Implement Density-Based Spatial Clustering of Applications with Noise (DBSCAN) or OPTICS to identify clusters of varying shapes and sizes in the data.
- **Spectral Clustering:** Consider using Spectral Clustering for non-linearly separable data, as it can handle complex structures and capture intricate relationships between data points.

3. Optimize Clustering Hyperparameters:

- **Elbow Method Enhancement:** Develop a more sophisticated approach for determining the optimal number of clusters using the Elbow method, such as the Silhouette Score or Gap Statistics, to ensure robust clustering results.
- **Grid Search and Cross-Validation:** Perform hyperparameter tuning using grid search and cross-validation techniques to fine-tune clustering algorithms' parameters for better performance.

4. Validation and Evaluation:

- **Internal Validation Metrics:** Utilize internal validation metrics like Silhouette Score, Davies-Bouldin Index, and Dunn Index to evaluate the quality of the clusters and compare different clustering algorithms.
- **External Validation Metrics**: If possible, gather external information, such as customer segments created by domain experts, to validate the clustering results' real-world relevance.

5. Visualizations and Interpretations:

- **Interactive Visualizations:** Create interactive visualizations using libraries like Plotly or Bokeh to allow users to explore and interact with the clusters and customer segments dynamically.
- **Feature Importance Analysis:** Conduct a feature importance analysis to identify which features significantly contribute to the separation of clusters. This will help in understanding the key factors influencing customer behavior.

6. Customer Profiling and Business Recommendations:

- **Customer Profiling:** Develop detailed customer profiles for each cluster, including demographic information, spending behavior, and preferences. This will enable businesses to tailor their marketing strategies and product offerings for each customer segment.
- **Business Recommendations:** Provide actionable business recommendations based on the insights gained from the customer segmentation. For example, suggest targeted marketing campaigns, personalized promotions, and inventory optimization strategies.

7. Online Implementation:

- **Real-time Segmentation:** Implement an online version of the customer segmentation model that can update customer clusters in real-time as new data becomes available. This will enable businesses to adapt quickly to changing customer behavior and market trends.
- **Integration with Business Systems:** Integrate the segmentation model with existing business systems, such as Customer Relationship Management (CRM) software or E-commerce platforms, to ensure seamless utilization of customer segments in day-to-day operations.

8. Ethical Considerations:

- **Privacy and Data Protection:** Ensure compliance with data privacy regulations and ethical considerations when collecting and using customer data for segmentation. Implement data anonymization techniques if needed.

9. Continuous Improvement:

- **Feedback Mechanism:** Establish a feedback mechanism to collect feedback from stakeholders and users regarding the effectiveness and usefulness of the segmentation model. Use this feedback to continuously improve the model and update business strategies.

Conclusion: Implementing the future work outlined above will enhance the Wholesale Customer Segmentation project's accuracy and utility, leading to more valuable insights and better business decisions. By continually refining the segmentation model and incorporating it into business operations, companies can improve customer satisfaction, increase sales, and achieve a competitive advantage in the market.

Concept Explanation :

Once upon a time, in the magical world of Data Science, there was a village called "Wholesale Kingdom." The villagers of this kingdom were busy running their businesses and selling various products to customers from different regions. But one day, they realized that they needed some magic to understand their customers better and improve their sales strategies.

So, they summoned the wise Data Scientist, who came with a shiny wand and a bag full of magical algorithms! The Data Scientist promised to help the villagers segment their customers into different groups so they could tailor their products and services accordingly.

Now, the first magic trick the Data Scientist used was called "Data Preprocessing." The Data Scientist cleaned and transformed the data to make it ready for analysis. They turned the data into a format that all the magical algorithms could understand.

Next, it was time for the most exciting part - the "Clustering Spell!" The Data Scientist used a special spell called "K-Means" to group the customers into clusters based on their similarities. Imagine grouping customers who love dragons together or those who adore unicorns!

The K-Means spell worked like this: The Data Scientist randomly placed some "Centroids" (special magical points) in the kingdom. Each centroid represented a cluster. Then, the Data Scientist assigned each customer to the nearest centroid based on their buying behaviors. It was like sorting customers into different houses based on their favorite magical creatures!

But the villagers were curious and asked, "How do we know how many clusters to create?" The Data Scientist smiled and said, "Fear not! We shall use the 'Elbow Method' to find the perfect number of clusters." They explained that the Elbow Method bends the graph like a wizard's staff, and the bending point suggests the optimal number of clusters.

With the clusters in place, the villagers were overjoyed to see how their customers magically sorted into different groups! They could now understand each group's preferences and design special potions and spells to cater to their needs.

But the adventure didn't end there! The Data Scientist decided to go even further and try other magical spells like "Density-Based Clustering" and "Spectral Clustering." These spells worked differently, but they were equally powerful in creating clusters that could capture unique customer behaviors.

As the Data Scientist explored more, they stumbled upon the "PCA Secret Scroll." This secret scroll allowed them to reduce the data's dimensions and visualize the clusters in a magical 2D world! Imagine seeing the clusters floating in the air like colorful balloons!

But wait, there's more! The Data Scientist also used the "Hierarchical Clustering Spell." This spell created a magical tree-like structure that connected the clusters in a hugging, dendrogram family!

After a series of magical adventures, the Data Scientist successfully helped the villagers unlock the secrets of their customers. The Wholesale Kingdom was now thriving, with each cluster of customers receiving personalized attention and tailored magical products.

And so, the tale of Wholesale Customer Segmentation came to a delightful end. The villagers lived happily ever after, using the magic of Data Science to make their customers smile and their businesses flourish.

So, dear readers, the moral of the story is that with the right Data Scientist and magical algorithms, anything is possible! Embrace the power of Data Science, and you too can unlock the hidden treasures hidden within your data. Happy Data Explorations! 🧙‍♀️✨

Exercise Questions :

1. What is the purpose of the Wholesale Customer Segmentation project?

Answer: The purpose of the Wholesale Customer Segmentation project is to analyze customer data from the Wholesale Kingdom and group customers into clusters based on their buying behaviors. This segmentation will help the villagers tailor their products and services to better meet the needs of each customer group.

2. How did the Data Scientist preprocess the data in this project?

Answer: The Data Scientist performed data preprocessing by checking for null values and data types, converting categorical variables to meaningful data, scaling the numeric features, and handling outliers. They also transformed the data to make it follow a more normal distribution.

3. What is the K-Means algorithm, and how does it work in this project?

Answer: K-Means is a clustering algorithm that groups data points into K clusters based on their similarities. In this project, the Data Scientist used K-Means to assign customers to clusters based on their buying behaviors. The algorithm randomly places centroids (representing clusters) and iteratively assigns customers to the nearest centroid until convergence.

4. How did the Data Scientist determine the optimal number of clusters to use in the K-Means algorithm?

Answer: The Data Scientist used the Elbow Method to find the optimal number of clusters. They plotted the within-cluster sum of squares for different numbers of clusters and looked for the "elbow" point, which indicates the point where adding more clusters doesn't significantly decrease the sum of squares.

5. What other clustering algorithms did the Data Scientist explore in this project?

Answer: The Data Scientist explored Density-Based Clustering and Spectral Clustering as alternative algorithms for customer segmentation. These algorithms have different

approaches to identifying clusters but are equally powerful in capturing unique customer behaviors.

6. How did the Data Scientist visualize the clusters in the magical 2D world?

Answer: The Data Scientist used the PCA (Principal Component Analysis) technique to reduce the data's dimensions to two components. These two components represent the most significant variations in the data. Then, they plotted the customers' data points in this 2D space, with each point colored according to its cluster.

7. What is the purpose of the Hierarchical Clustering Spell?

Answer: The Hierarchical Clustering Spell creates a tree-like structure called a dendrogram that connects the clusters. It helps visualize how clusters are related to each other and allows for hierarchical grouping of data points.

8. How did the Data Scientist handle outliers in the data?

Answer: The Data Scientist used the Interquartile Range (IQR) method to handle outliers. They identified the upper and lower bounds for each feature based on the IQR and replaced any data points outside these bounds with the respective bound value.

9. What benefits does customer segmentation provide to businesses in the Wholesale Kingdom?

Answer: Customer segmentation allows businesses in the Wholesale Kingdom to understand each customer group's preferences and needs better. This understanding enables them to tailor their products, marketing strategies, and customer service to maximize customer satisfaction and overall business performance.

10. How can the Wholesale Kingdom continue using the insights gained from this project to improve their business strategies?

Answer: The Wholesale Kingdom can continue monitoring customer behavior within each cluster and track changes over time. They can use this information to make data-driven decisions, launch targeted marketing campaigns, optimize inventory management, and introduce new products that align with each cluster's preferences. Additionally, the Wholesale Kingdom can periodically re-run the customer segmentation

analysis to adapt to evolving customer demands and ensure their business remains successful and enchanting!