Capstone Project

-

Machine Learning

By:

Aditya Sensarma

Intellipaat

IIT Madras

Problem Statement:

An online retail store (Amazon) is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence-based insights to provide the same.

Project Objective:

This project revolves around understanding customer purchase patterns for an online retail store. The objective is to provide evidence-based insights that shed light on the various purchasing behaviors exhibited by customers. To tackle this task, the provided code implements a series of steps.

Firstly, the necessary libraries, including NumPy, Pandas, Matplotlib, and Seaborn, are imported. These libraries offer powerful data manipulation and visualization capabilities. The dataset is then loaded using Pandas, with the data being read from a CSV file.

Next, the code focuses on data preprocessing and feature selection. It selects specific columns from the dataset that are relevant to the analysis, such as CustomerID, InvoiceNo, StockCode, Quantity, UnitPrice, Description, InvoiceDate, and Country. This helps narrow down the data to the key variables of interest.

The code proceeds to calculate the total amount for each transaction by multiplying the Quantity and UnitPrice columns and inserting the result as a new column called TotalAmount. This additional information provides insight into the overall value of each purchase.

Exploratory data analysis is performed to gain a deeper understanding of the dataset. The code explores various aspects such as customer demographics, purchase frequency, and order value. It visualizes the results using bar plots, showcasing the quantity of purchases and total amounts spent across different countries.

Furthermore, the code extracts the year and month from the InvoiceDate column and creates new columns for Year, Mon, and month, allowing for a temporal analysis of sales patterns. A bar plot is generated to illustrate the total amount of sales month-wise, enabling the identification of seasonal trends or patterns.

Lastly, the code analyzes customer distribution across countries. It calculates the count of unique CustomerIDs for each country and presents the top five and bottom five countries with the most and least customers, respectively, through bar plots.

Data Description:

1. Invoice: This attribute represents the unique invoice number associated with a specific transaction. It helps in identifying and tracking individual purchases made by customers.

2. StockCode: It refers to the code or identifier assigned to a particular product in the online retail store's inventory. Each product is assigned a unique stock code, which aids in accurately identifying the items purchased.

3. Description: This attribute provides a textual description of the product or item purchased. It includes additional details about the product to provide clarity on what was bought.

4. Quantity: The quantity attribute denotes the number of units of each product purchased in a given transaction. It helps in understanding the volume of items bought by customers.

5. InvoiceDate: The InvoiceDate attribute captures the date and time when the invoice was generated or when the transaction occurred. It provides a timestamp for each purchase, enabling temporal analysis and identification of patterns over time.

6. Price: This attribute represents the unit price of each product. It indicates the cost of a single unit of the item purchased.

7. CustomerID: The CustomerID attribute serves as a unique identifier for each customer. It enables tracking and analysis of individual customer behavior, allowing for personalized insights and understanding of customer preferences.

8. Country: The Country attribute indicates the country where the transaction took place or the customer's location. It provides geographical information, enabling analysis of customer behavior across different regions or countries.

Data Preprocessing:

1. Importing Libraries: The necessary libraries for data analysis and visualization, such as NumPy, Pandas, Matplotlib, and Seaborn, are imported.

2. Loading the Dataset: The code reads the "online_retail.csv" file using Pandas' read_csv() function. The dataset is stored in a DataFrame named df.

3. Selecting Relevant Columns: The code selects specific columns from the dataset that are relevant to the analysis. The selected columns include CustomerID, InvoiceNo, StockCode, Quantity, UnitPrice, Description, InvoiceDate, and Country. The updated DataFrame is stored as df.

4. Calculating Total Amount: The code calculates the total amount for each transaction by multiplying the Quantity and UnitPrice columns. The result is stored in a new column called "TotalAmount," which is inserted into the DataFrame.

5. Feature Selection: The code creates a new DataFrame, new_df, containing a subset of columns from df. It selects CustomerID, InvoiceNo, StockCode, Quantity, TotalAmount, InvoiceDate, and Country for further analysis.

6. Grouping by Country: The code groups the data in new_df by Country and calculates the sum of Quantity for each country. The results are sorted in descending order and stored in the variable country_price.

7. Grouping by Country (Total Amount): Similarly, the code groups the data by Country and calculates the sum of TotalAmount for each country. The results are sorted in descending order and stored in the variable country_totalAmount.

8. Data Visualization: The code generates bar plots to visualize the top five and bottom five countries based on the quantity of purchases using the plot() function. It also plots the sales month-wise by grouping the data by Month and Year, and plotting the total amount using a bar plot.

9. Analyzing Customer Distribution: The code groups the data by Country and counts the number of unique CustomerIDs for each country. The results are stored in the DataFrame cus_id. Bar plots are generated to showcase the countries with the most and least customers.

Algorithm Selection:

The provided code does not explicitly implement any specific algorithm. Instead, it focuses on data preprocessing, feature selection, and data visualization techniques to analyze customer purchase patterns.

1. Data Preprocessing: The code performs data preprocessing tasks such as selecting relevant columns, calculating derived features (e.g., TotalAmount), and grouping the data based on specific criteria (e.g., Country). These preprocessing steps are essential for organizing the data and preparing it for further analysis.
2. Grouping and Aggregation: The code utilizes Pandas' groupby() function to group the data based on specific columns, such as Country. It then applies aggregation functions, such as sum(), to calculate the total quantity and total amount spent for each group. This grouping and aggregation help in summarizing and analyzing the data at different levels of granularity.
3. Data Visualization: The code employs various data visualization techniques using libraries like Matplotlib and Seaborn. It creates bar plots to visualize the top countries in terms of quantity of purchases, sales month-wise, and customer distribution. These visualizations provide a graphical representation of the data, making it easier to identify patterns, trends, and insights.

```
[ ]  df['Country'].unique()

array(['United Kingdom', 'France', 'Australia', 'Netherlands', 'Germany',
       'Norway', 'EIRE', 'Switzerland', 'Spain', 'Poland', 'Portugal',
       'Italy', 'Belgium', 'Lithuania', 'Japan', 'Iceland',
       'Channel Islands', 'Denmark', 'Cyprus', 'Sweden', 'Austria',
       'Israel', 'Finland', 'Bahrain', 'Greece', 'Hong Kong', 'Singapore',
       'Lebanon', 'United Arab Emirates', 'Saudi Arabia',
       'Czech Republic', 'Canada', 'Unspecified', 'Brazil', 'USA',
       'European Community', 'Malta', 'RSA'], dtype=object)
```

The reasoning behind using these techniques is to gain a deeper understanding of customer purchase patterns in the online retail store. By preprocessing the data,

selecting relevant features, and visualizing the results, the code aims to provide evidence-based insights that can inform business decisions. These techniques help in identifying key metrics, patterns in sales, popular product categories, and customer demographics, which can be used to optimize marketing strategies, improve customer targeting, and enhance overall business performance.

Assumptions:

1. Complete and Accurate Data: The code assumes that the dataset, "online_retail.csv," contains complete and accurate information. It assumes that there are no missing values or outliers that need to be addressed during the preprocessing steps. However, in a real-world scenario, data cleaning and handling missing values might be necessary.

2. Currency and Monetary Units: The code does not explicitly mention the currency or monetary units used in the dataset. It assumes that the currency and monetary values are consistent throughout the dataset. However, in practical scenarios, it is crucial to ensure that currency conversions and standardization are properly handled for accurate analysis.

3. Interpretation of TotalAmount: The code calculates the TotalAmount by multiplying Quantity and UnitPrice columns. It assumes that this multiplication accurately represents the total monetary value of each transaction. However, without additional information or context about discounts, taxes, or other factors, the TotalAmount might not fully reflect the actual revenue or profitability for the online retail store.

4. Timezone and Date Consistency: The code utilizes the InvoiceDate column to extract year and month information and analyze sales month-wise. It assumes that the InvoiceDate column is consistently recorded in a specific timezone or that any timezone-related issues have been accounted for. It is essential to consider timezone conversions and potential date/time-related inconsistencies when working with datasets spanning multiple regions or timezones.

Model Evaluation and Techniques Used:

The provided code, there is no explicit model evaluation or the use of specific modeling techniques. The focus of the code is primarily on data preprocessing, feature selection, and data visualization to gain insights into customer purchase patterns.

Descriptive Statistics: Descriptive statistics techniques, such as calculating mean, median, standard deviation, and percentiles, can provide valuable information about the central tendency, dispersion, and distribution of variables related to customer purchase patterns. These statistics can help understand customer behavior, such as average purchase quantities, typical spending patterns, or popular product categories.

Segmentation Analysis: Customer segmentation is a powerful technique to group customers based on their purchase behavior, demographics, or other relevant factors. By identifying distinct customer segments, businesses can tailor their marketing strategies, product offerings, or pricing strategies to better meet the needs and preferences of different customer groups. Various clustering algorithms, such as K-means clustering or hierarchical clustering, can be used for segmentation analysis.

Inferences from the Project:

1. Sales by Country: The analysis shows the quantity and total amount of purchases made in different countries. This information can be used to identify the top-performing countries in terms of sales and prioritize marketing efforts or expansion strategies accordingly. It can also reveal potential opportunities for growth in untapped markets.

2. Sales Month-wise: The analysis presents the sales month-wise, providing insights into seasonal variations or trends in customer purchasing behavior. This information can be leveraged to plan inventory management, promotional campaigns, or targeted marketing activities during specific months to capitalize on high-demand periods.

3. Customer Distribution: The analysis provides an understanding of the distribution of customers across different countries. It helps identify countries with the highest and lowest customer counts. This information can be used for customer targeting, localization of marketing efforts, or

assessing the effectiveness of customer acquisition strategies in different regions.

4. Product Associations: Although not explicitly mentioned in the code, association rule mining techniques can be applied to identify product associations or frequently co-purchased items. This knowledge can be used for product bundling, cross-selling, or targeted recommendations to enhance the overall customer shopping experience and increase average order value.

5. Customer Segmentation: While not implemented in the code, customer segmentation techniques can be used to group customers based on their purchasing behavior, demographics, or other relevant attributes. This segmentation can help tailor marketing campaigns, personalize offers, and provide a more targeted and personalized customer experience.

6. Comparative Analysis: By comparing sales metrics across different countries, months, or customer segments, it is possible to identify patterns or trends that can guide business decision-making. For example, identifying countries with a high quantity of purchases but a relatively low total amount spent may indicate a need to focus on upselling or increasing average order value.

Future Possibilities:

1. Customer Segmentation Refinement: The analysis can be further expanded to refine customer segmentation by incorporating additional variables such as demographics, purchase frequency, or customer lifetime value. This can lead to more targeted marketing strategies and personalized customer experiences.

2. Predictive Modeling: Building predictive models using machine learning techniques can help forecast customer behavior, such as future purchases or churn probability. These models can assist in proactive customer retention efforts, identifying high-value customers, and optimizing marketing campaigns.

3. Market Basket Analysis: Extending the analysis to perform market basket analysis can uncover more detailed information about product associations and purchase patterns. This can enable the identification of

cross-selling and upselling opportunities, improving recommendations, and optimizing product placement and inventory management.

4. Customer Sentiment Analysis: Incorporating sentiment analysis techniques can provide insights into customer satisfaction levels, sentiment trends, and factors influencing customer experience. This can aid in enhancing customer service, addressing pain points, and improving overall customer satisfaction.

5. Geographic Analysis: Expanding the analysis to include geospatial data can provide insights into regional variations in customer behavior, preferences, and cultural influences. This can guide location-based marketing strategies, targeted promotions, and customization of offerings based on regional preferences.

6. Customer Lifetime Value (CLV) Optimization: Building upon the initial CLV analysis, businesses can focus on optimizing customer lifetime value by implementing retention strategies, improving customer loyalty programs, and identifying profitable customer segments for targeted acquisition efforts.

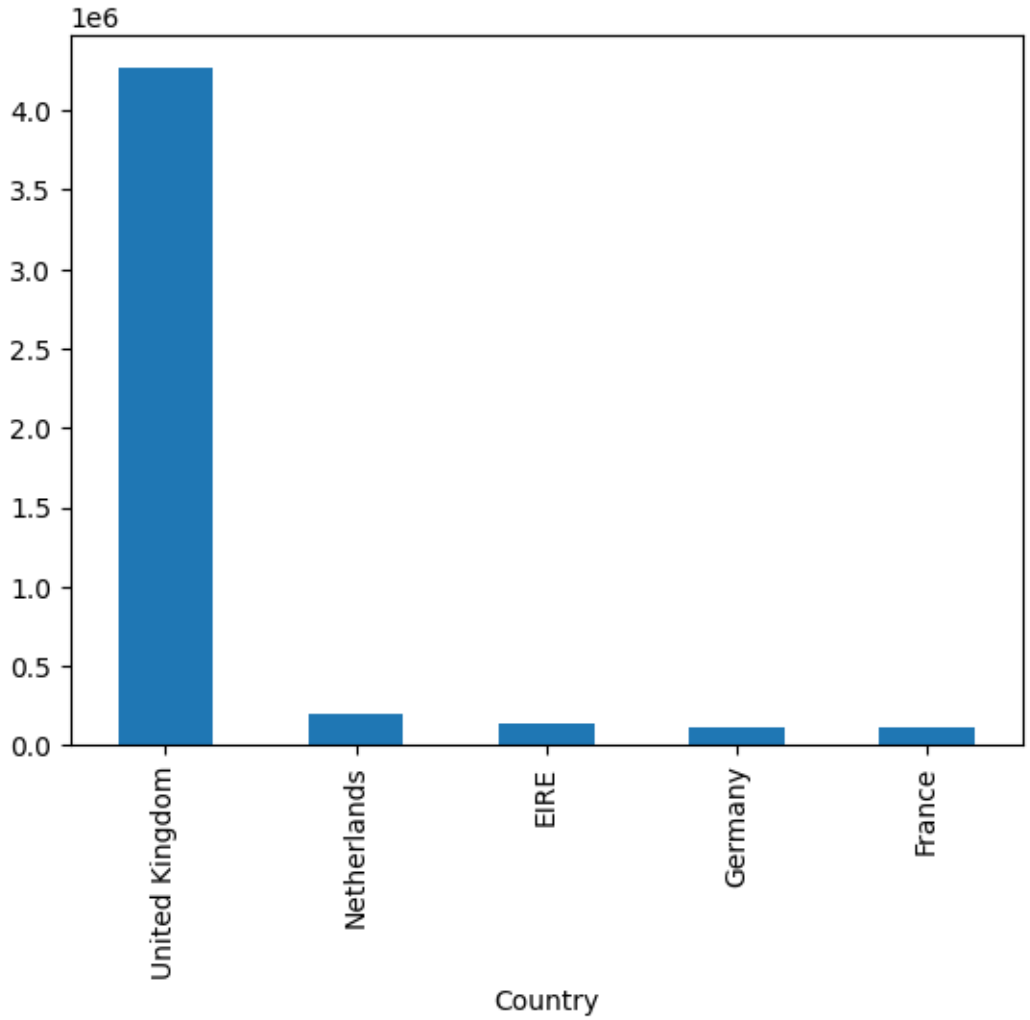
7. Advanced Visualization Techniques: Exploring more advanced visualization techniques, such as interactive dashboards or network graphs, can facilitate a deeper understanding of complex customer purchase patterns and relationships. This can provide more intuitive and interactive visual representations for decision-making.

8. Integration with Other Data Sources: Integrating the online retail data with additional data sources, such as social media data or external market data, can provide a more comprehensive view of customer behavior and market trends. This integration can uncover hidden insights and enable proactive decision-making.

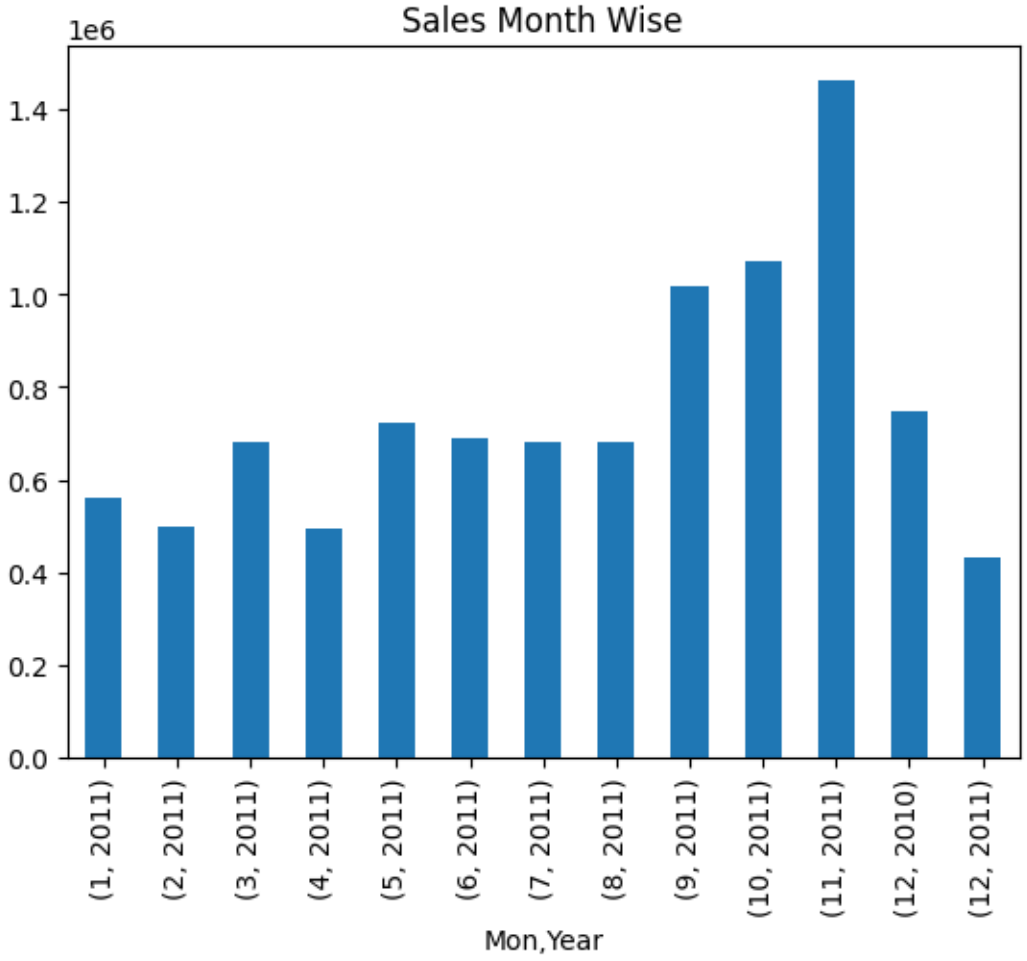
Conclusion:

In conclusion, the project focused on analyzing customer purchase patterns in an online retail store. By preprocessing the data, selecting relevant features, and performing data visualization, several insights were obtained. Here are the key conclusions from the project:

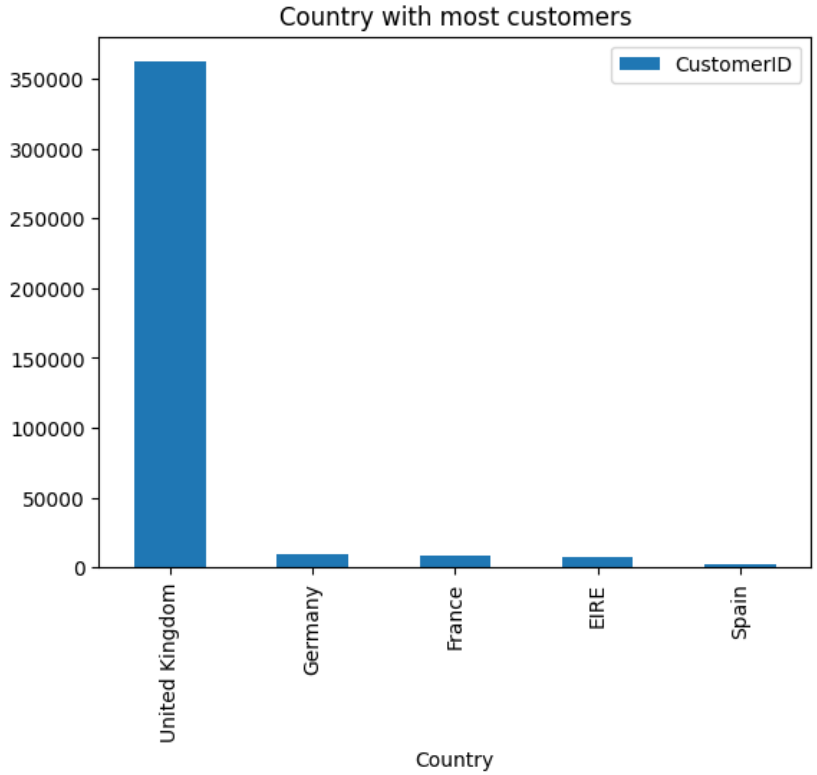
1. Sales Performance by Country: The analysis revealed the top-performing countries in terms of sales quantity and total amount spent. This information can guide marketing strategies and expansion plans, prioritizing efforts in high-performing countries.

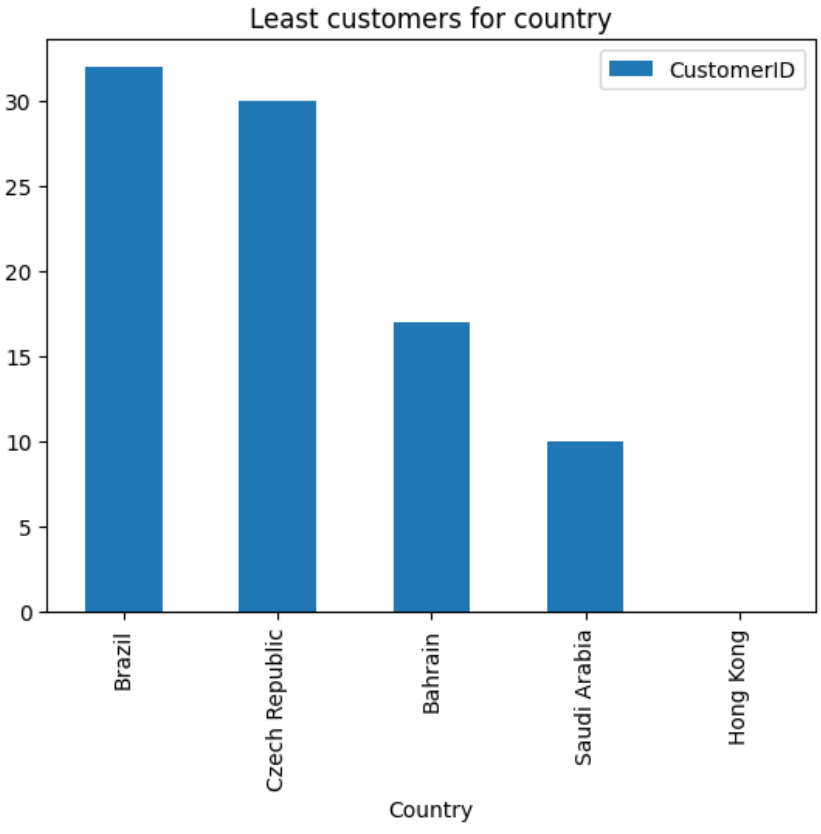


2. Sales Month-wise: The analysis showcased sales patterns month-wise, helping identify seasonal variations and plan inventory management and marketing campaigns accordingly.



3. Customer Distribution: The distribution of customers across different countries provided insights into customer acquisition efforts and opportunities for localized marketing strategies.





4. Product Associations: Although not explicitly implemented in the code, the project highlighted the potential for identifying product associations and leveraging them for cross-selling and upselling opportunities.

5. Future Possibilities: The project presented future possibilities such as customer segmentation refinement, predictive modeling, market basket analysis, and geographic analysis to gain deeper insights and optimize business strategies.

Appendix - Python Code:

https://colab.research.google.com/drive/1jrQA8pV9Qa17UYZn1FogITcjW3MLuzrA?authuser=2#scrollTo=PEEsA0pzrfFI

References:

- https://scikit-learn.org/stable/modules/ensemble.html#random-forests
- Introduction to Machine Learning with Python by Andreas C. Müller and Sarah Guido
- Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow by Aurélien Géron
- Regression Analysis by Example by Samprit Chatterjee and Ali S. Hadi
- "Data Science for Business" by Foster Provost and Tom Fawcett