Analysis of Online Retail Sales Performance

1

Teammates and Roles

Team :

1.Ananyaa Matta - AnanyaaMatta@my.unt.edu - 11707045

2.Vamsi krishna kanchukommala - vamsikrishnakanchukommala@my.unt.edu - 11666316

3.Ashish Reddy rampurm - ashishreddyrampurm@my.unt.edu - 11594372

4.Sumalika Puppala-sumalikapuppala@my.unt.edu – 11597319

Roles :

Collecting and analyzing various datasets – opensource datasets: Vamsi Krishna

Data Visualization – Sumalika

Data Analysis – Ananyaa

Statistical Testing – Ashish

Documentation and reporting – Vamsi Krsihna, Sumalika, Ananyaa, Ashish

2

Abstract

Online shopping is a growing industry that provides a wealth of data that can be used to gain important insights into the dynamics of sales performance. In order to identify the complex patterns and variables impacting sales performance, this study explores the thorough analysis of sales data collected from online retail platforms. This study aims to offer a comprehensive understanding of the dynamics influencing sales in the digital marketplace by utilizing sophisticated analytical approaches such as clustering algorithms, regression analysis, and time series forecasting.

Some of the insights this project seeks to derive are, identification of clients who are the main sources of revenue is one important factor that is being examined. This aims to identify the primary income drivers and provides insight into their preferences and purchase behaviour through extensive analysis. These kinds of information not only help companies attract and keep valuable customers, but they also help them develop focused marketing campaigns that maximise profits.

3

Additionally, this project closely studies how various nations operate in the internet retail space. It clarifies the fundamental causes affecting sales variances across various geographic regions by identifying the best and worst performing nations. These kinds of data are very helpful in developing customized plans for entering new markets, distributing resources efficiently, and realizing unrealized potential in emerging areas.

In addition, the project provides year-round sales trend analysis, revealing seasonal variations and demand trends. Businesses can maximize resource allocation, marketing campaigns, inventory management, and other strategies to take advantage of profitable opportunities and reduce risks by recognizing general trends and peak seasons.

Furthermore, the study examines the correlation between the quantity of products acquired and earnings, utilizing thorough hypothesis testing techniques to determine its importance. Businesses may make well-informed decisions and implement effective pricing plans by measuring the impact of purchase volume on revenue creation. This provides significant insights into customer behaviour and spending habits.

4

Finally, the project explores the identification of significant average sales variations among nations, explaining variations in customer preferences, purchasing power, and market dynamics. These insights give companies a sophisticated grasp of the worldwide market.

For businesses navigating the ever-changing world of e-commerce, the analysis of online retail performance is essentially an essential component, providing actionable insights that improve operational efficiency, drive strategic decision-making, and promote sustainable growth in a market that is becoming more and more competitive.

5

DATASET: https://www.kaggle.com/datasets/tunguz/online-retail

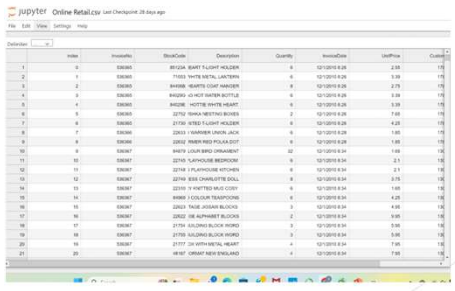
Description: This dataset includes data for an Analysis on an online retail store that ships to 38 countries. The dataset contains information on purchases for a year. Currency used for payment is in US dollars,selling various products, including gifts, household items, and accessories. The data includes information about customers, products, quantities sold, unit prices, and transaction dates.

Features: The dataset typically includes features such as Customer ID, Invoice Number, Stock Code, Quantity, Product Description, Quantity, Unit Price, and Invoice Date, Country

Format: The data is usually provided in a structured format such as CSV (Comma-Separated Values)

We have downloaded the dataset from the Kaggle. Once downloaded, we have preprocesed and analyzed the data using programming languages like Python and various data analysis tools and techniques to derive insights into online retail sales performance.

6



7

Exploratory Data Analysis

We have many data visualization techniques to show different relationships between different coloumns in the dataset some of them that we plotted are:

- Customers who brought in the most revenue
- The Top 5 revenue generating countries.
- The Least 5 revenue generating countries.
- Sales Trend throughout the year
- Relationship between number of items purchased and revenue, with a hypothesis test to prove it's significance
- Significant mean difference in sales between countries etc

8

Statistics Used:

Hypothesis Testing:

Hypothesis testing is a fundamental statistical technique used to make inferences about a population parameter based on sample data. In the context of analyzing retail sales performance, hypothesis testing can be used to assess the significance of observed differences or relationships in sales metrics. To determine statistical significance, to find relationship between number of items purchased and revenue, with a hypothesis test to prove it's significance. To draw conclusions about population parameters, assess the null hypothesis with the and data.

T-testing: The t-test is used to determine if there is a significant difference between the means of two independent groups. Here in this retail sales analysis project, we can use the t-test to compare sales performance metrics between two distinct groups.

For example: Compare the average sales (e.g., total revenue, quantity sold) between two customer segments, such as new customers versus returning customers.

9

Mann-Whitney U test:

The Mann-Whitney U test is a non-parametric test used to determine if there is a significant difference between the distributions of two independent groups.

In the retail sales analysis project, the Mann-Whitney U test can be applied when the assumptions of the t-test are not met (e.g., non-normal distribution of data). For example: Compare sales performance metrics between different product categories

Pearson Correlation: Pearson correlation is a statistical method used to measure the strength and direction of the linear relationship between two continuous variables. In the context of analyzing retail sales performance, statistic=0.91, pvalue=0.0

The correlation coefficient = 0.9, which is a positive correlation. The p-value = 0.0 which is less than 0.05.We reject the null hypothesis. There exist a statistically significant relationship between these two variables

10

- So, In this project, we have used hypothesis test to find significant difference in revenue between United Kingdom and the other countries and there's a correlation between Quantity of items sold and Revenue generated
- We must need to run a two-sample t-test to determine if the difference in mean sales between the UK and second most contributing country, Ireland is statistically significant.
- We have continued to check the assumptions such as , if the samples are independent, if the sales for one country should not affect the sales for the other country and if the samples are normally distributed: We checked this by visualizing the distribution of the sales data for each country using a histogram or a Q-Q plot.
- If these assumptions are not met, then there would be a need to use a different statistical test such as a non-parametric test like the Mann-Whitney U test.
- Then, we did Mann-Whitney U test, There is a difference in mean sales between UK and Ireland

Where we got the, Significance level = 0.05

- We reject the null hypothesis since the p-value is less than the significance level.
- Therefore, we accept the alternate hypothesis that there is a difference in mean sales between these two countries

11

Resources:

Online Courses:

Coursera: "Data Analysis and Visualization" or "Data Science Specialization"

edX: "Data Science MicroMasters" or "Statistics and Data Science MicroMasters"

Books:

"Python for Data Analysis" by Wes McKinney

"R for Data Science" by Hadley Wickham and Garrett Grolemund

"Data Science for Business" by Foster Provost and Tom Fawcett

Online Platforms:

Kaggle: Explore datasets, participate in competitions, and access tutorials on data analysis and machine learning.

DataCamp: Interactive courses covering data analysis, visualization, and machine learning using Python and R.

12

Related Projects:

Customer Segmentation and Personalization: Use clustering algorithms to divide up your clientele according to their tastes, demographics, and purchase patterns. In order to improve consumer happiness and loyalty, use these categories to customize pricing tactics, product recommendations, and marketing efforts.

Sales and Demand Prediction: To project future trends in sales and demand, create time series forecasting models. Optimize inventory management, resource allocation, and production planning for businesses by including elements like seasonality, promotions, and external events into your projections to increase their accuracy.

Analysis of Market Baskets and Cross-Selling Techniques: Use association rule mining tools to investigate relationships between products that are frequently purchased together. To raise average order value and maximize income, find cross-selling possibilities and optimize product placements, promotions, and bundling methods.

13

DATA SPECIFICATION:

We used structured dataset. Dataset has total has 541909 rows and 7 columns
Columns in the dataset:

1. Invoice Number: Unique identifier for each transaction.
2. Stock Code: Stock code represents a reference code assigned to each product
3. Product Description: Description of the product.
4. Quantity: Number of units of a particular product purchased in a transaction
5. Invoice Date: Invoice date denotes the date and time when a transaction took place
6. Unit Price: The price of a single unit of the product
5. Customer ID: Unique identifier for each customer.
7. Country: Name of the country

14

Predictions and Results:

These are all the things that we have analyzed from the dataset we have taken using different statistical methods and exploratory data analysis

-Most purchases made falls around 3 dollars.

-In January,sales was 618,239 which took a dip and fluctuated till July and then steadily rose from August to November where it peaked at 1,349,975 dollars and it reduced in December by a margin of 17,000 dollar. The actual sales in December was 1,332,949 dollars. The least revenue were recorded in February and April with 456,164 and 464,301 dollars respectively

-Revenue recorded for the first quarter was 1,707,078 dollars and it had a steady rise till the 4th quarter where it peaked at 3,697,855 dollars.

-Customers without ID brought the largest revenue of 1,662,045 dollars. The customer that brought in the highest revenue is customer ID '18102' with 258,518 dollars

-There is a positive relationship between quantity of items purchased and revenue and has a statistical significance to it,proven by the hypothesis test. Revenue increases as more items are purchased. This does not mean a causal effect.

-Customers from the United Kingdom brought in the most revenue at about 8,012,072 dollars. The second country is Ireland with 251,272 dollars.

-Countries that generated the most revenue also bought the most items this goes to prove our hypothesis of positive relation between quantity of items purchased and revenue.

-Saudi Arabia generated the least revenue among all the countries. This store made only 138 dollars from Saudi Arabia.

15

Design

Technologies Used: Python(Pandas, Numpy, Seaborn, SciPy, scikit-learn)

Project Scope Definition:

- Specifing the project's goals, which should include actionable insight extraction and analysis of sales performance using online retail databases.
- Describing the methods and instruments that will be applied. For example, Python can be used for data visualization and analysis.

Data Collection and Cleaning:

- Collecting data on internet retail transactions from reputable sources or artificial intelligence datasets.
- Completing data cleaning activities, such as dealing with outliers, duplicates, and missing numbers.
- Making sure the data is valid and intact before analyzing it.

Exploratory Data Analysis (EDA):

- To comprehend the properties and structure of the dataset, perform an early EDA..
- Calculating summary statistics, examine correlations between variables, and visualize distributions.
- Determining the trends, patterns, and possible causes that could be affecting sales performance

16

Statistical Analysis and Hypothesis Testing:

- Statistical analysis, such as price elasticity and correlation analysis, can be used to investigate relationships between variables.
- To evaluate statistical significance, develop hypotheses about factors that affect sales performance and run hypothesis tests (such as t-tests and ANOVA).
- Analyze results and offer data-driven suggestions for improvement.

Documentation and Reporting:

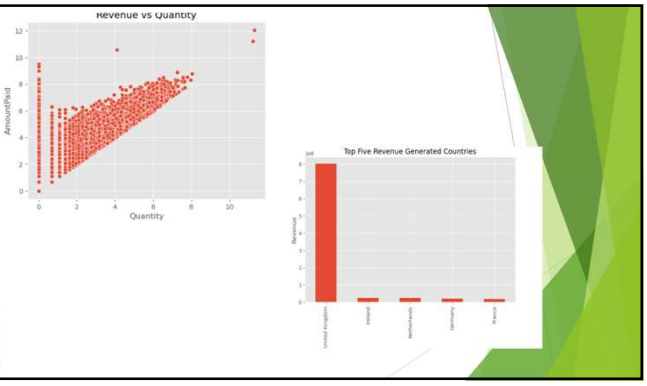
- Document the methodology, findings, and insights obtained throughout the project.
- Prepare a comprehensive report and presentation

17

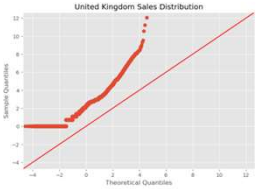
VISUALIZATIONS:

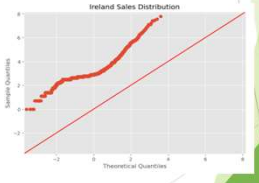
- Some of the visualizations from the code output

18

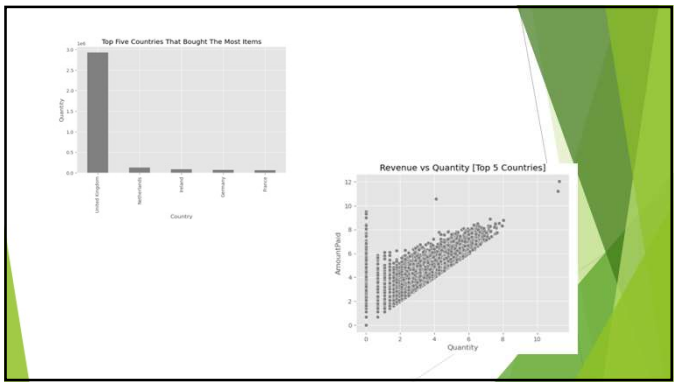


19





20



21

Future Enhancements:

- -Launching a global campaign to appeal to customers in countries across the world.
- -For an unbiased analysis on customers, the store's database should be optimized to capture the identity of all customers so we can further get insights into the customers who truly bring in more revenue. This way, they would not be left out of target discounts and loyalty programs.This could help prevent customer churn.
- -This store could target their advertisements towards customers from the least performing countries to attract them to purchase more. There could be other strategies like subsidising shipping fees for these customers.
- -If feasible, the store could run periodic discount sales across the year to attract new customers and retain old customers.
- -To curb the issue of customers largely purchasing items within 3 dollars,the store could maximize their advertisement on products that cost a little more. For a better reach, the stake-holders cloud build a recommender system that recommends items to customers based on their activities on the site.
- -The United Kingdom generates the largest chunk of revenue, to retain customers from this country,there could be periodic discounts or loyalty programs targeted at these customers.
- -To maximise customer satisfaction, there could be a system setup to track customers' feedback.

22

THANK YOU

23