

**ANALYSING THE PATTERN SALES BETWEEN THE PRODUCTS FOR ONLINE RETAIL DATASET**

**PROJECT REPORT**

***Submitted by***

**NEERAJ. B - E0320031**

***In partial fulfilment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

**(Artificial Intelligence and Data Analytics)**

**Sri Ramachandra Engineering and Technology**

**Sri Ramachandra Institute of Higher Education and Research, Porur, Chennai -600116**

**MARCH, 2022**



**ANALYSING THE PATTERN SALES BETWEEN THE PRODUCTS FOR ONLINE RETAIL DATASET**

**PROJECT REPORT**

***Submitted by***

**NEERAJ. B - E0320031**

***In partial fulfilment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

**(Artificial Intelligence and Data Analytics)**

**Sri Ramachandra Engineering and Technology**

**Sri Ramachandra Institute of Higher Education and Research, Porur, Chennai -600116**

**MARCH, 2022**



**BONAFIDE CERTIFICATE**

Certified that this project report **“ANALYSING THE PATTERN SALES BETWEEN THE PRODUCTS FOR ONLINE RETAIL DATASET”** is the bonafide work of **NEERAJ. B - E0320031** who carried out the internship work under my supervision.

**Signature of Faculty Mentor Signature of Vice-Principal**

|  |  |
| --- | --- |
| **Prof. Ashokkumar P**  Assistant Professor  Sri Ramachandra Engineering and Technology  Porur  Chennai-600116 | **Prof. M. Prema**  Vice-Principal  Sri Ramachandra Engineering and Technology  Porur  Chennai-600116 |

**Evaluation Date:**

|  |  |
| --- | --- |
| **TABLE OF CONTENTS**  **Title** | **Page** |
| 1. Domain Introduction   * 1. Data Analytics   2. Objective   3. Problem Statement | 7  7  7  7 |
| 2. Project Description  2.1 Data Visualization  2.2 Association Rule Mining  2.3 Apriori Algorithm | 8  8  8  8 |
| 3.Literature Survey | 9 |
| 4. Methodology  4.1 Development Steps [Workflow] | 10  10 |
| 5. Components Used  5.1 Python  5.2 System Requirement | 11  11  11 |
| 6. Technology Used   6.1 Jupyter Notebook | 12  12 |
| 7. Implementation  7.1 Code and Output Screenshots | 13  13 |
| 8. Result  8.1 Conclusion  8.2 Future Scope | 26  26  26 |
| 9. References | 27 |
| 10. Worklog | 28 |
| 11. Offer Letter | 29 |

**ACKNOWLEDGEMENT**

I express my sincere gratitude to our Chancellor, Vice-Chancellor and our sincere gratitude to our Provost **Dr. V.Raju** and our Vice-Principal **Prof. M.Prema** for their support and for providing the required facilities for carrying out this study.

I wish to thank my faculty supervisor(s), **Prof. Ashokkumar P,** Department of Computer Science and Engineering, Sri Ramachandra faculty of Engineering and Technology and **Mr. Abhishek Kumar Jha, Co-founder of** **Utkarshini Edutech** for extending help and encouragement throughout the project. Without his/her continuous guidance and persistent help, this project would not have been a success for me.

I am grateful to all the members of Sri Ramachandra Faculty of Engineering and Technology, my beloved parents and friends for extending the support, who helped us to overcome obstacles in the study.

**ABSTRACT**

The project one accomplished during my internship at Utkarshini Edutech Pvt Lmtd during the time frame of Jan to Feb is in the field of data science, in which we are given a raw dataset, gather the data, clean the data, prepare the data, then finding the pattern between the products using Associative Mining.  Later the derived solution was used to discover the output using visualisation tools such as graphs, plots and various other tools which were implemented for the first project – Analysing the Pattern Sales Between The Products For Online Retail Dataset.

1. **DOMAIN INTRODUCTION**
   1. **Data Analytics**

* The science of analysing data using algorithms, statistics, and technology is known as data analytics. Data analytics focuses on using programmes, data, and computational tools to explore and discover relevant insights in large amounts of data.
  1. **Objective**

The Main objective is to find and Analyze the Pattern Sales between the Products for Online Retail dataset using Association Data Mining.

* To find Which country sold more products.
* To find What are the Top 10 selling products.
* To know Which month has highest sale for Germany.
* To Analyze the Total sale in weekday and Weekend for Non-UK.
* In which month The maximum shopping is done.
  1. **Problem Statement**

To analyze the association among products bought together by customers and also to find the patterns in customer behavior based on their history of transactions.

1. **PROJECT DESCRIPTION**
   1. **Data Visualization**

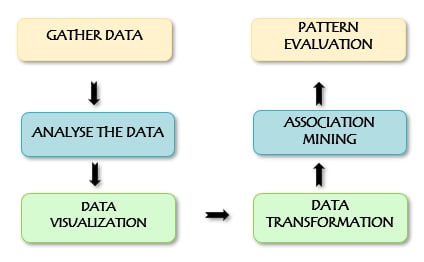
* The process of translating information into a visual representation, such as a map or graph, in order to make data easier to analyze and extract insights from is known as data visualization. The main goal of data visualization is to make discovering patterns, trends, and outliers in large data sets easier.
  1. **Association Rule Mining**
* The Association Rule is an unsupervised learning strategy that looks for the dependency of one data item on another and maps accordingly to make it more profitable.
* It looks for intriguing connections or links between the variables in the dataset.
  + Marketing, Basket Data Analysis (or Market Basket Analysis) in retailing, clustering, and classification are some of the uses of Association Rule Mining
  1. **Apriori Algorithm**
* The Apriori algorithm is a well-known data mining algorithm. It's used to find frequently occurring item sets as well as related association rules.
* The Apriori algorithm's main goal is to construct an association rule between different things. The association rule outlines the relationship between two or more items. Frequent pattern mining is another name for the Apriori algorithm.

1. **LITERATURE SURVEY**

* **Scalable algorithms for association mining:**
* **Link:**  https://sci-hub.hkvisa.net/10.1109/69.846291
* The finding of association rules has become a significant problem in knowledge discovery and data mining. We compare the new algorithms to the old ones in an experiment, and we find that the new algorithms outperform the old ones by more than an order of magnitude on our test databases.
* **Fast rule-based heart disease prediction using associative classification mining:**
* **Link:** <https://sci-hub.hkvisa.net/10.1109/IC4.2015.7375725>
* Associative Classification is a new and exciting technique that combines associative rule mining and classification. Many academics have been drawn to this technique because it produces accurate classifiers with effective rules. Associative classifiers are effective for applications that require the highest level of predicted accuracy. The healthcare business collects a vast amount of data that is not mined for hidden information that can be used to make decisions.
* **Comparative analysis of association rule mining algorithms:**
* **Link:** <https://sci-hub.hkvisa.net/10.1109/INVENTIVE.2016.7830203>
* The primary goal of this study is to evaluate the effectiveness of various association rule generation algorithms. Apriori, Eclat, Dclat, FP-growth, FIN, AprioriTID, Relim, and H-Mine are the algorithms considered in this comparison. Finally, we discovered that the DCLAT method outperformed other algorithms in this comparison analysis.

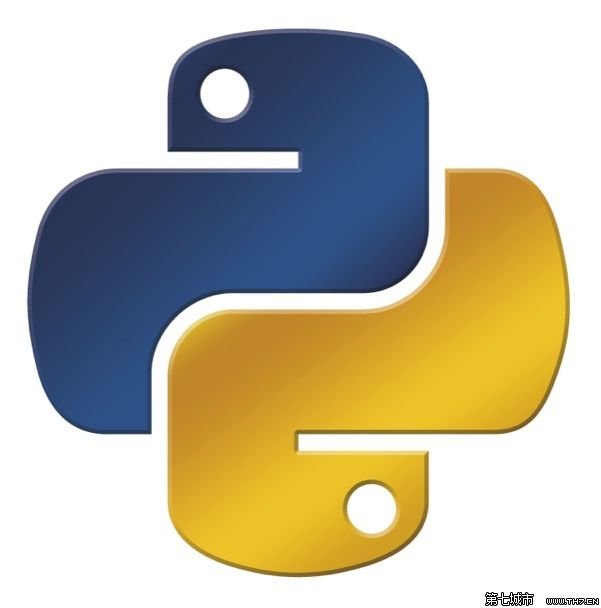
1. **METHODOLOGY**

**4.1. Development Steps**

* **Gather Data:** Kaggle was utilized to gather information about the subject.
* **Analyze the Data:** The data was studied using a variety of analytical methodologies such as correlation, shape, mean, and others to find hidden patterns and interactions between distinct elements.
* **Data Visualization:** The evaluated result was shown to make the hidden links and patterns in the data easier to understand.
* **Data Transformation:** Data transformation is the process of changing data from one format to another, usually from a source system's format to a destination system's needed format.
* **Association Mining:** The goal of association rule mining is to discover the rules that control how and why certain products/items are frequently purchased together.
* **Pattern Evaluation:** Determine the pattern's level of interest. To make data understandable to the user, summarization and visualization are used.

**5. COMPONENTS USED**

**5.1 Python**

* Python is a versatile programming language that can be used for a wide range of tasks. A straightforward and intuitive language that performs as well as its primary competitors.
* A simple and natural language that is just as powerful as the major contenders'.
* Because it is open source, anyone can contribute to its development. Code that is almost as comprehensible as plain English.
* Appropriate for running errands on a regular basis, given the short time it takes to improve.

**5.2 System Requirements**

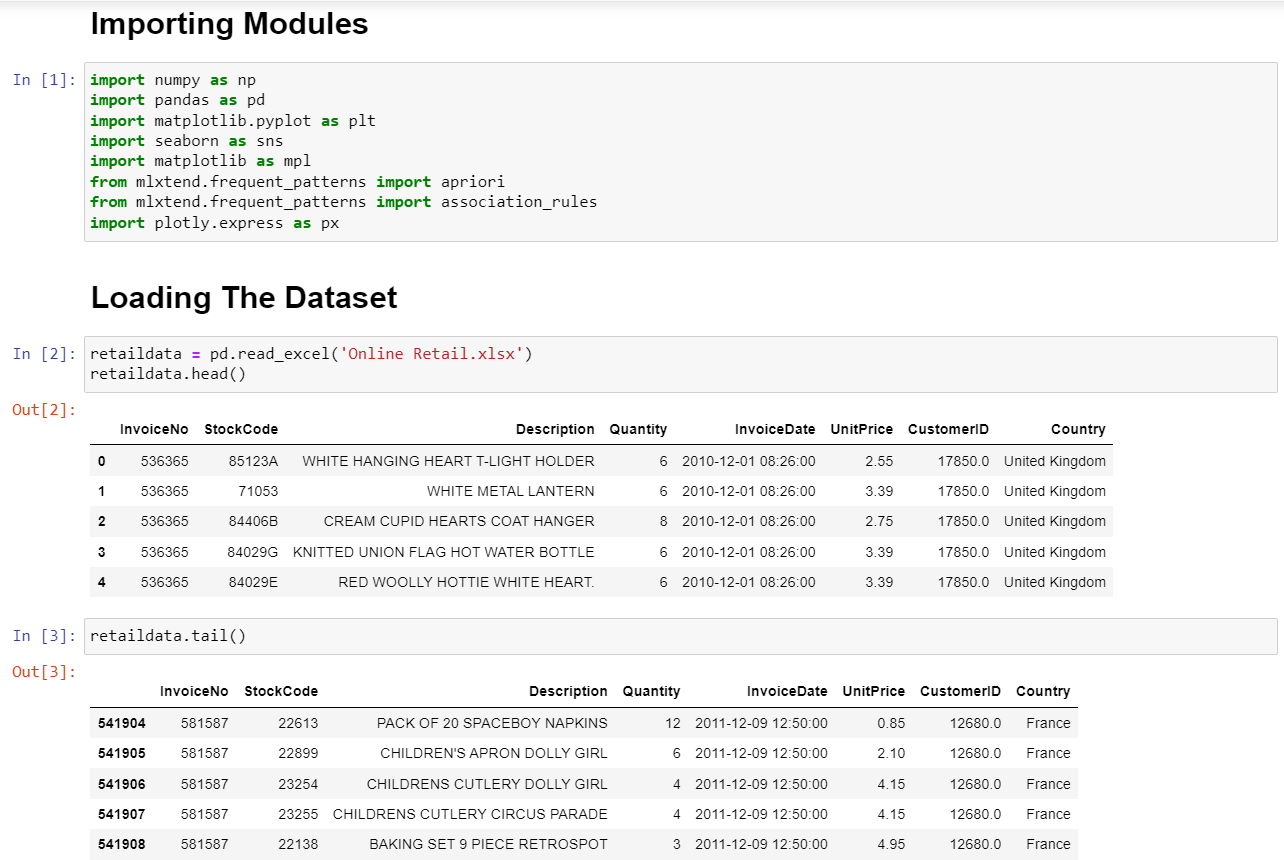
* Operating System: Windows 7 or later (64 bit).
* Disk Space: 1.32 GB (does not include disk space for IDE or tools).
* Tools: Windows PowerShell 5.0 or newer (it is pre-installed with windows 10)

1. **TECHNOLOGY USED**
   1. **Jupyter Notebook**

* ****Jupyter Notebook is a free, open-source web software that lets you create and share documents with live code, equations, visualisations, and narrative text. It can be used for a range of activities, including data cleansing and transformation, numerical simulation, statistical modelling, data visualisation, and machine learning.

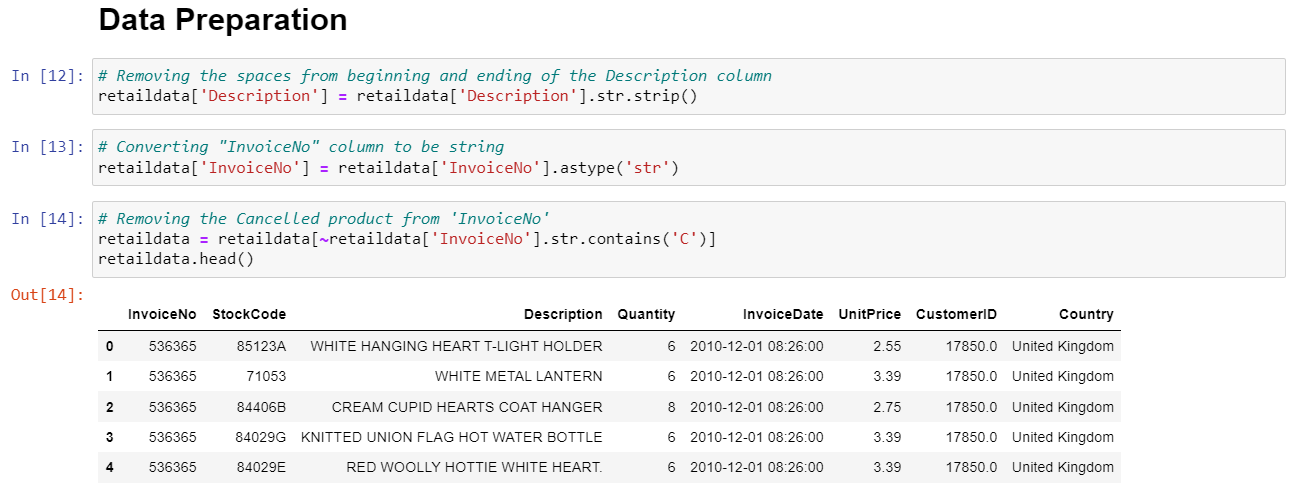
1. **IMPLEMENTATION**

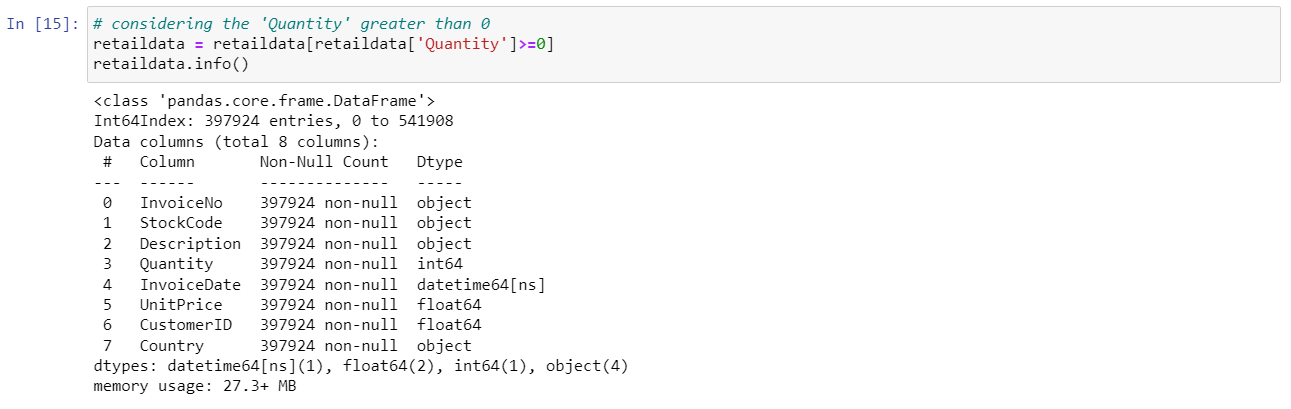
**7.1 Code and Screenshot**

****

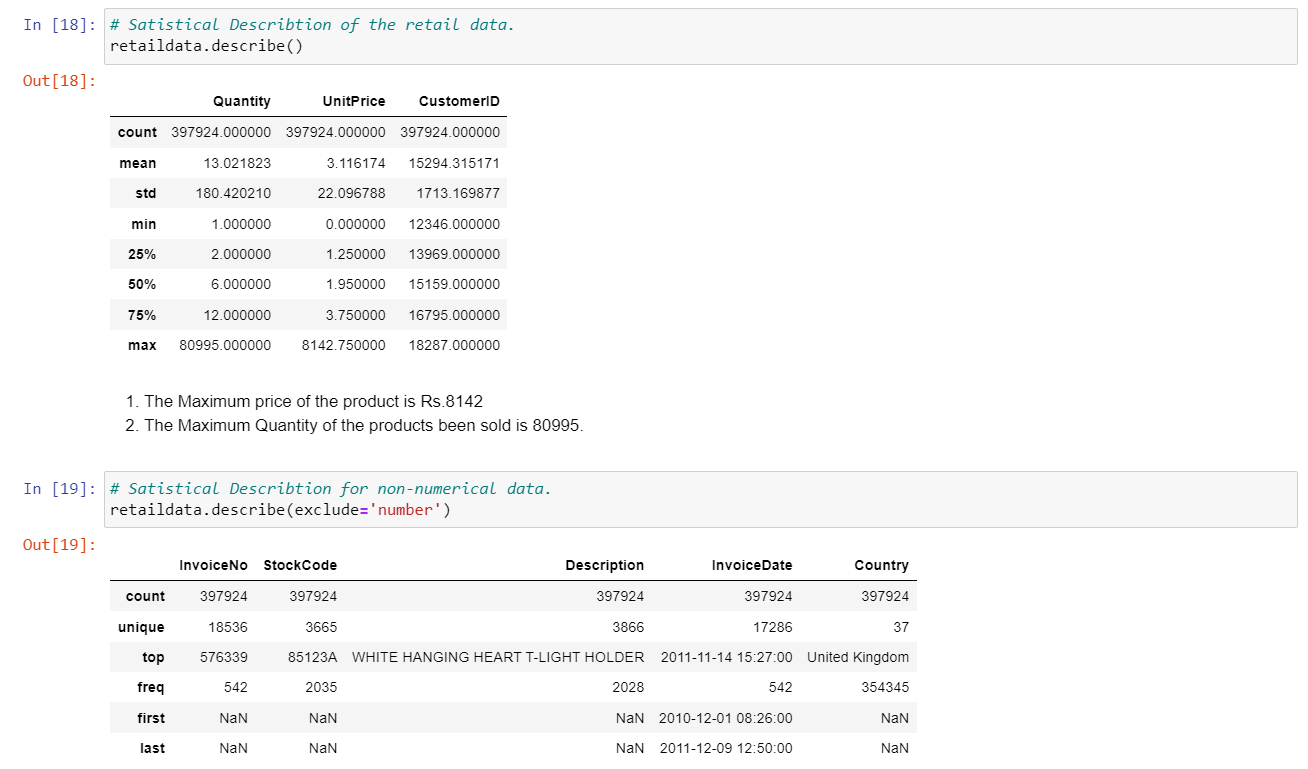
****

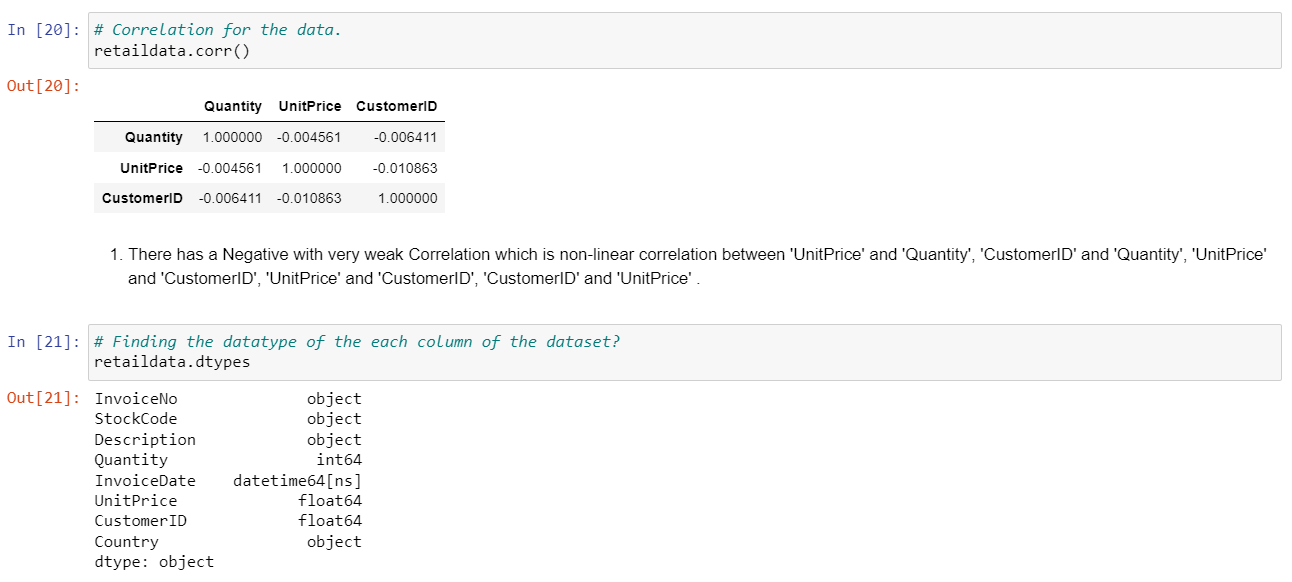
****

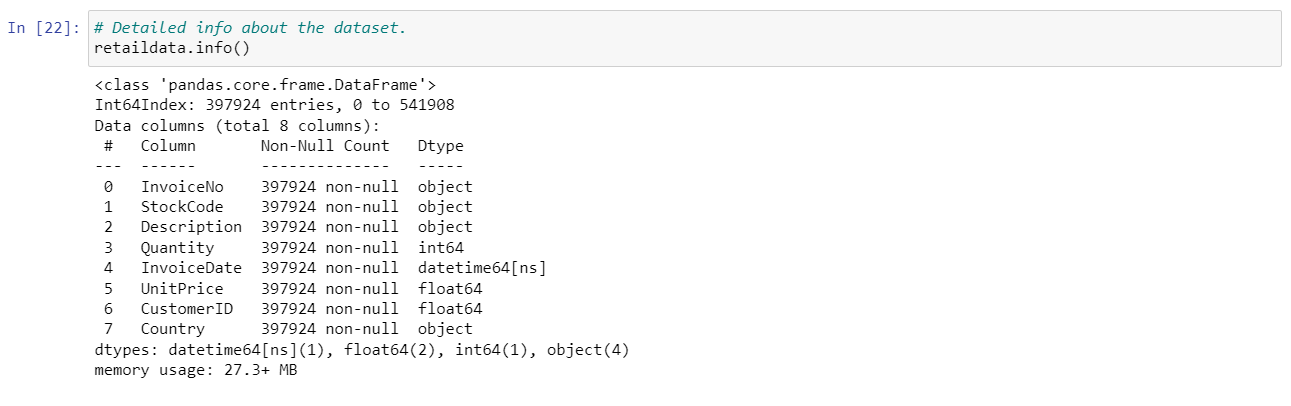
****

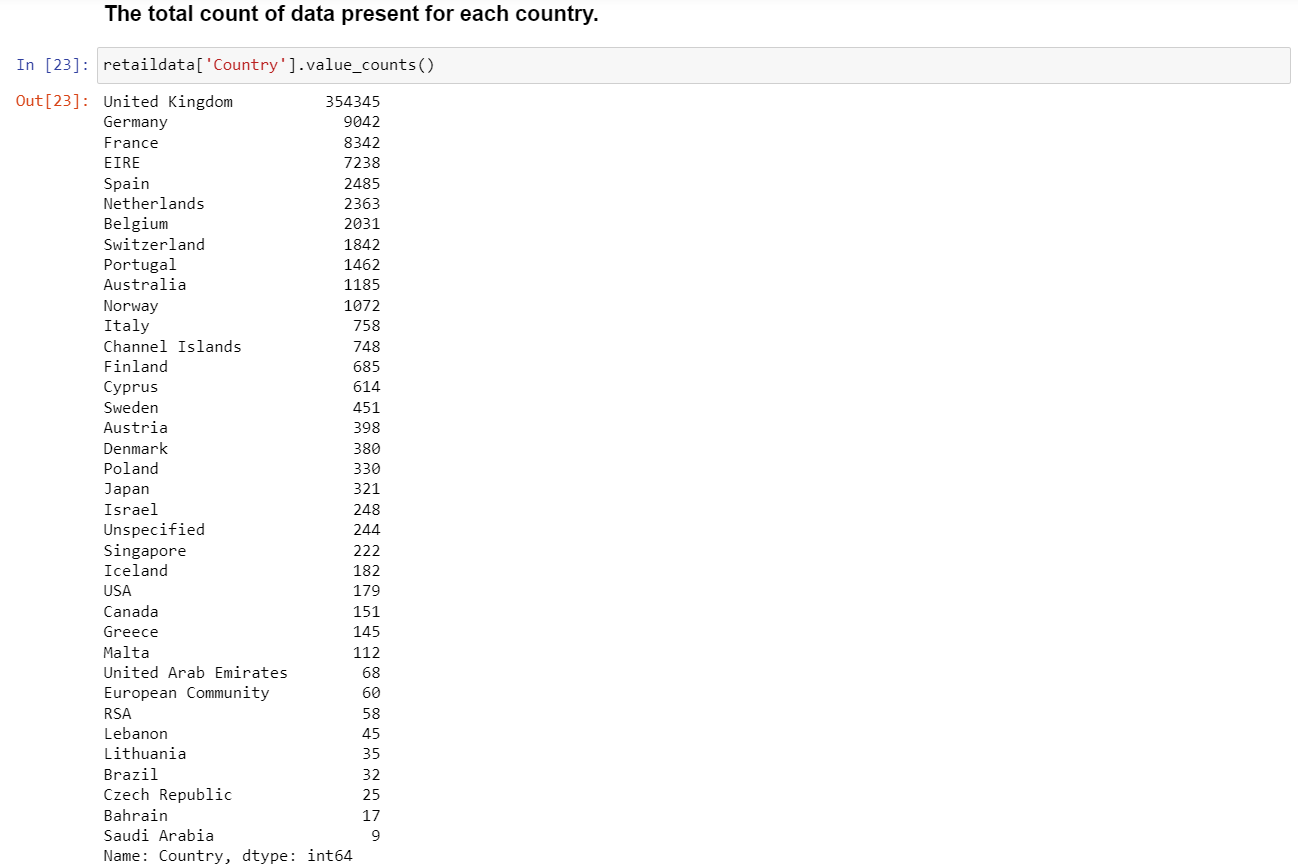
****

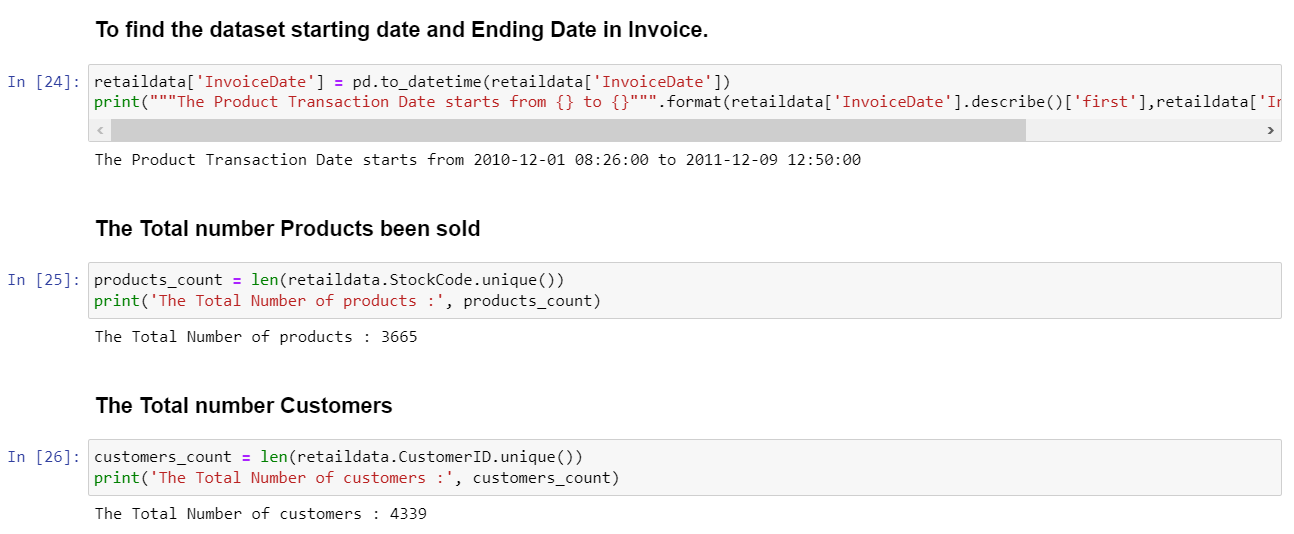


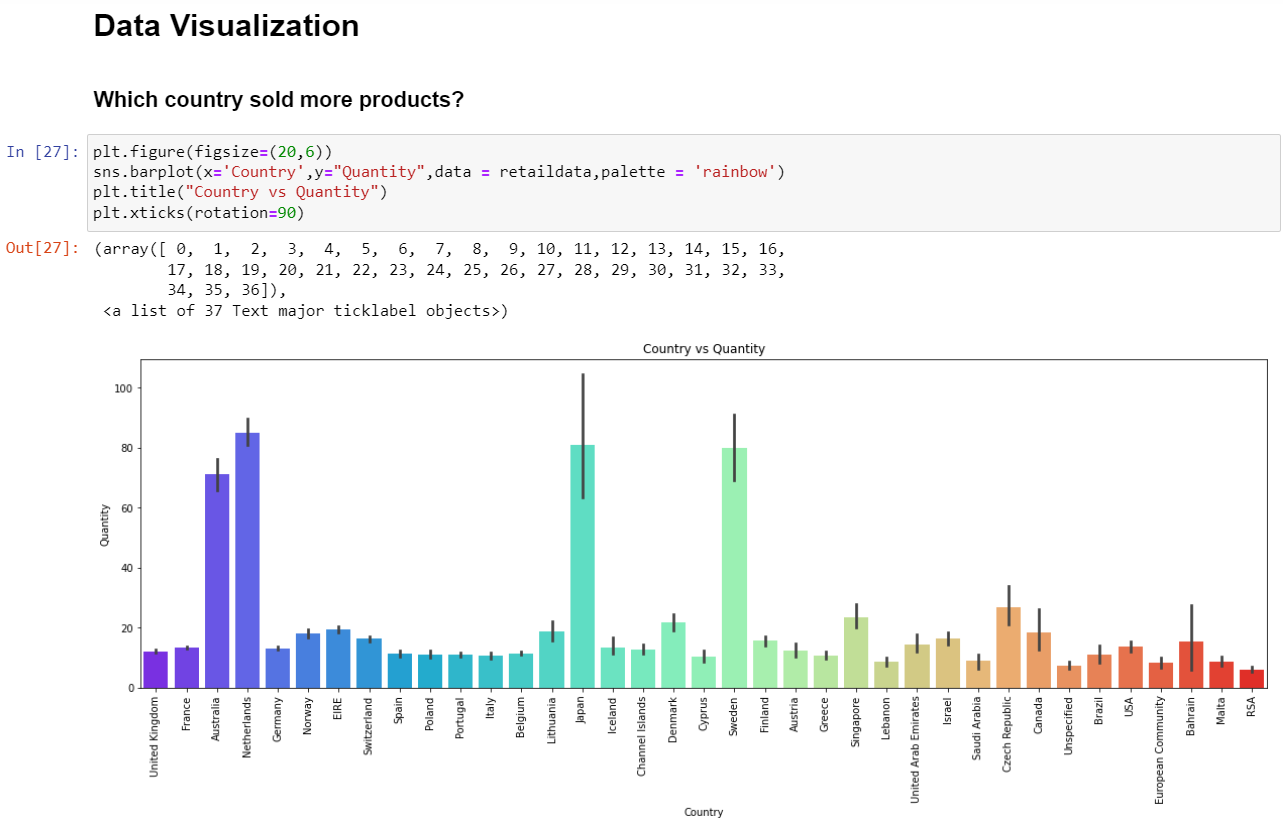


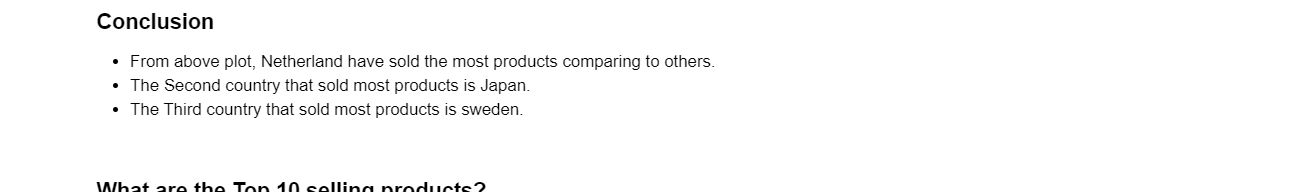


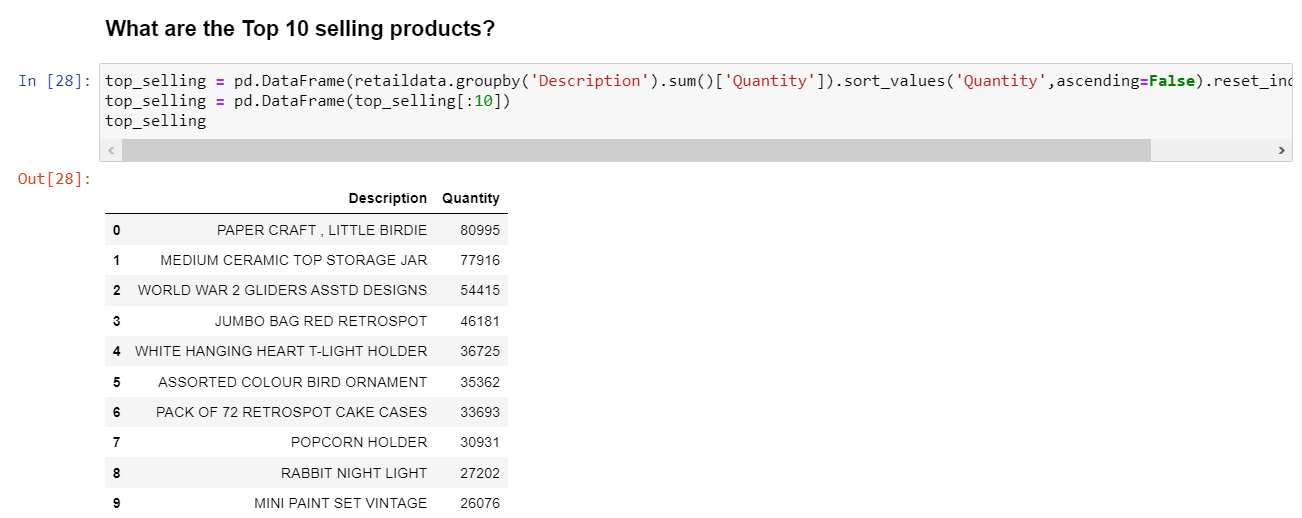


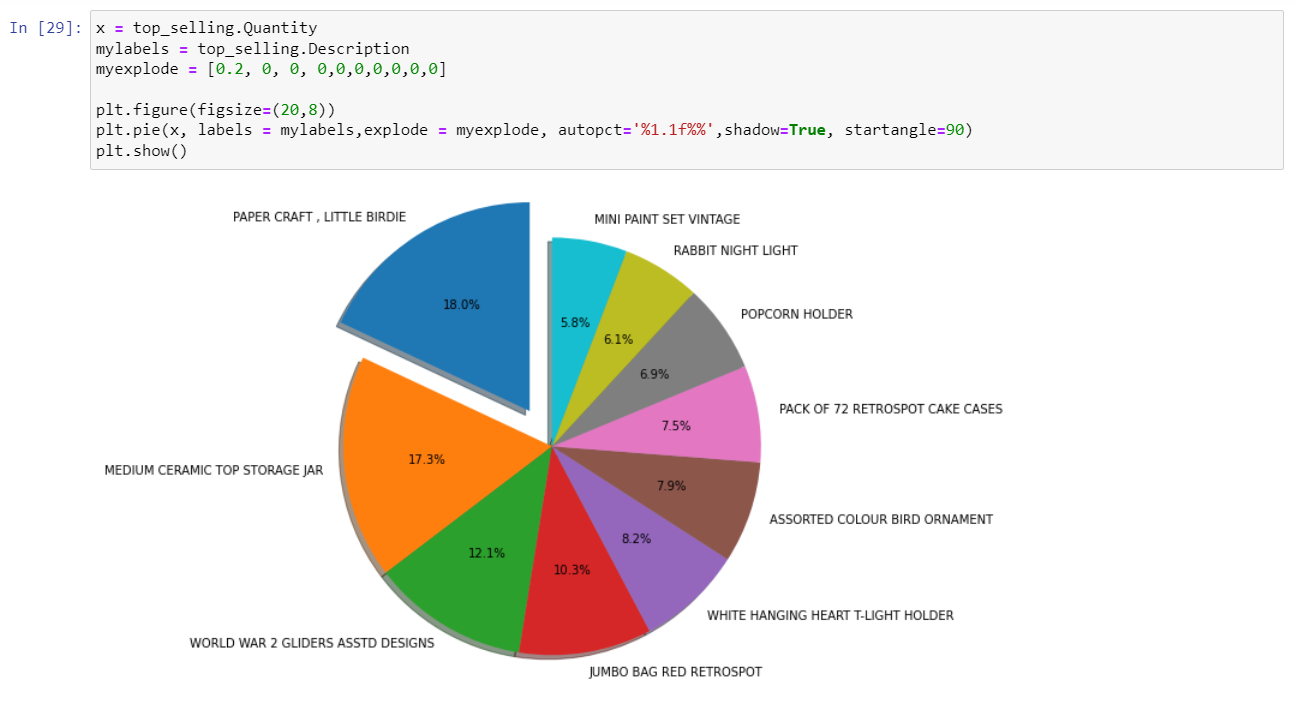


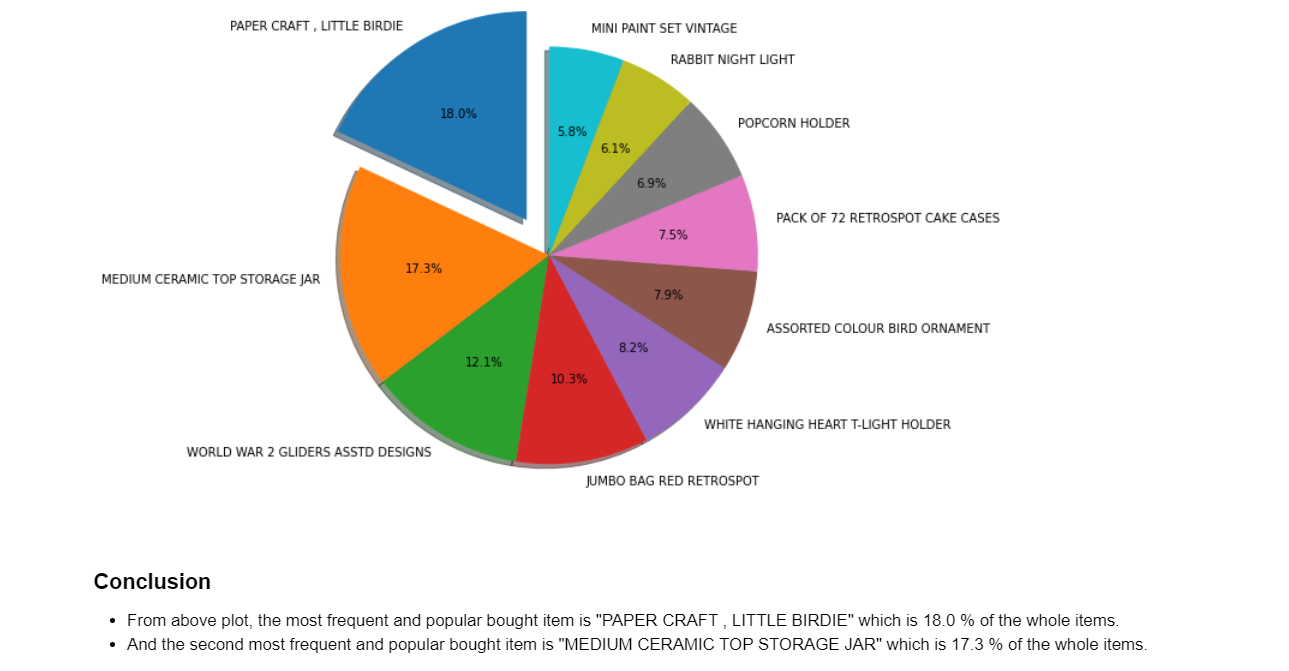




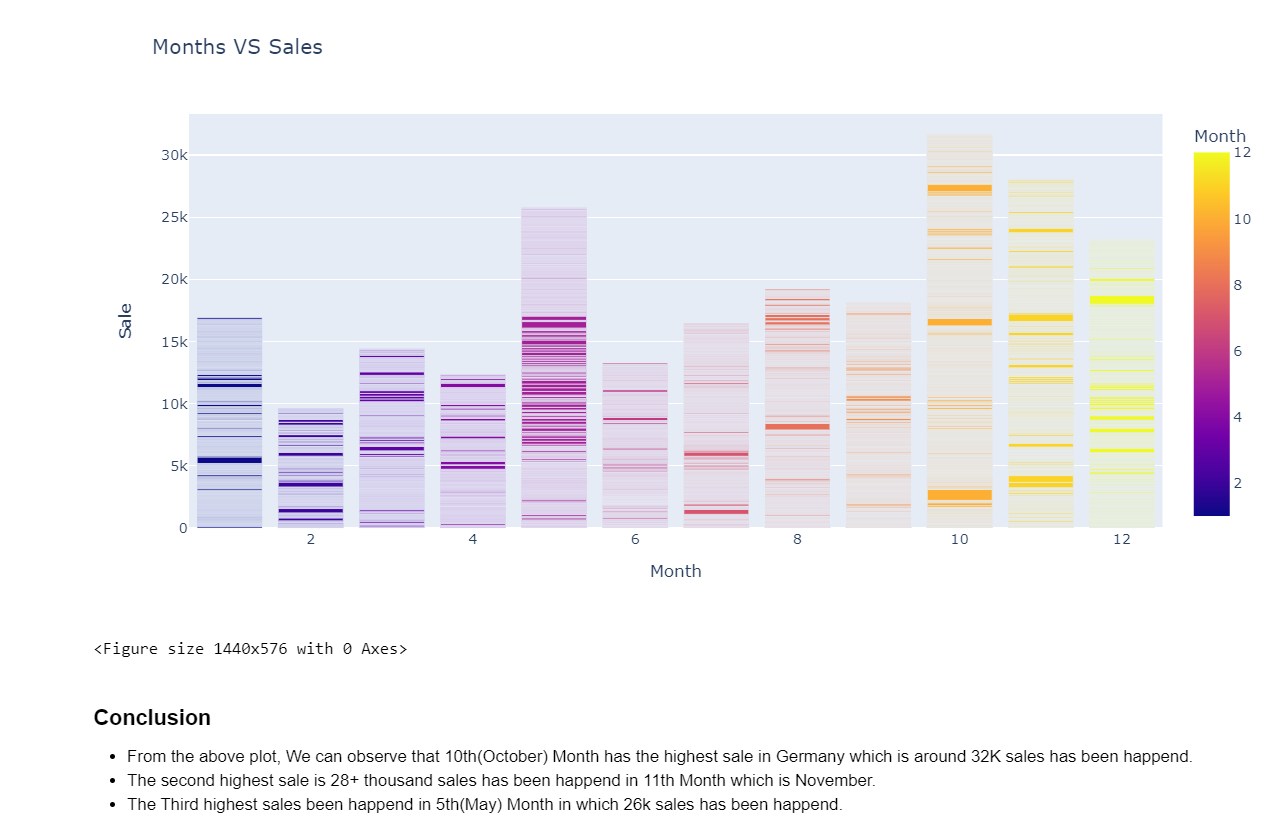


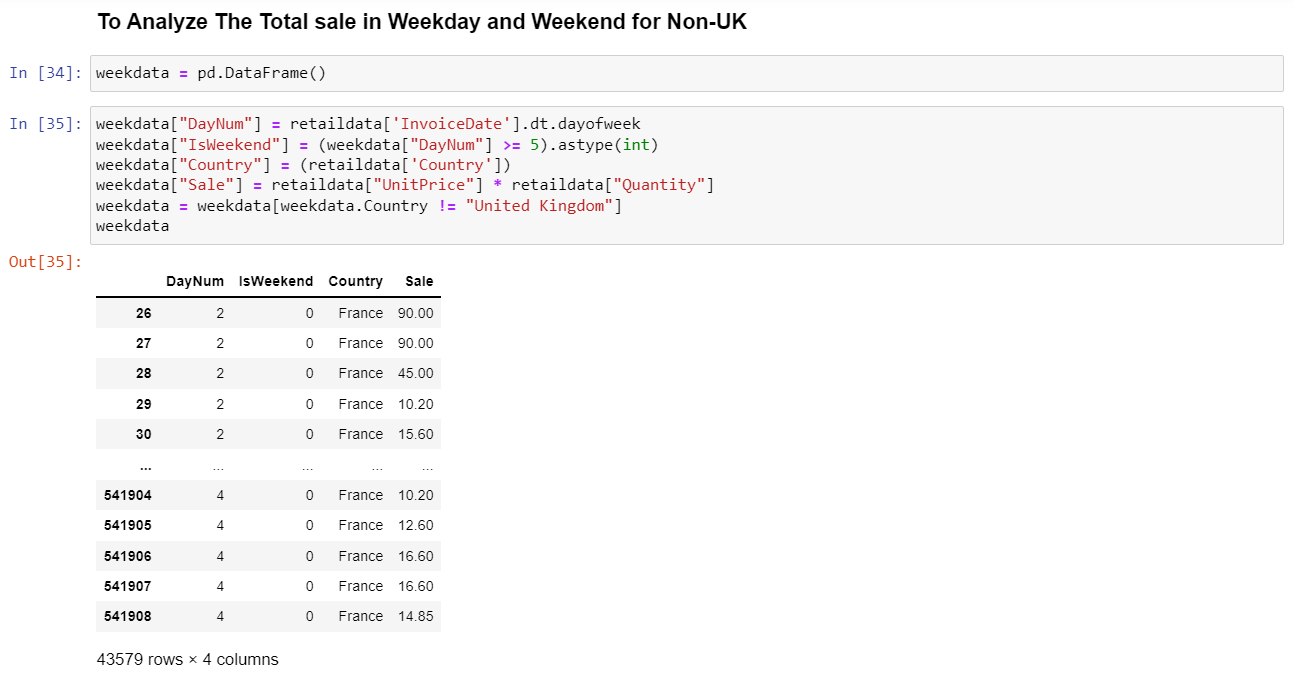


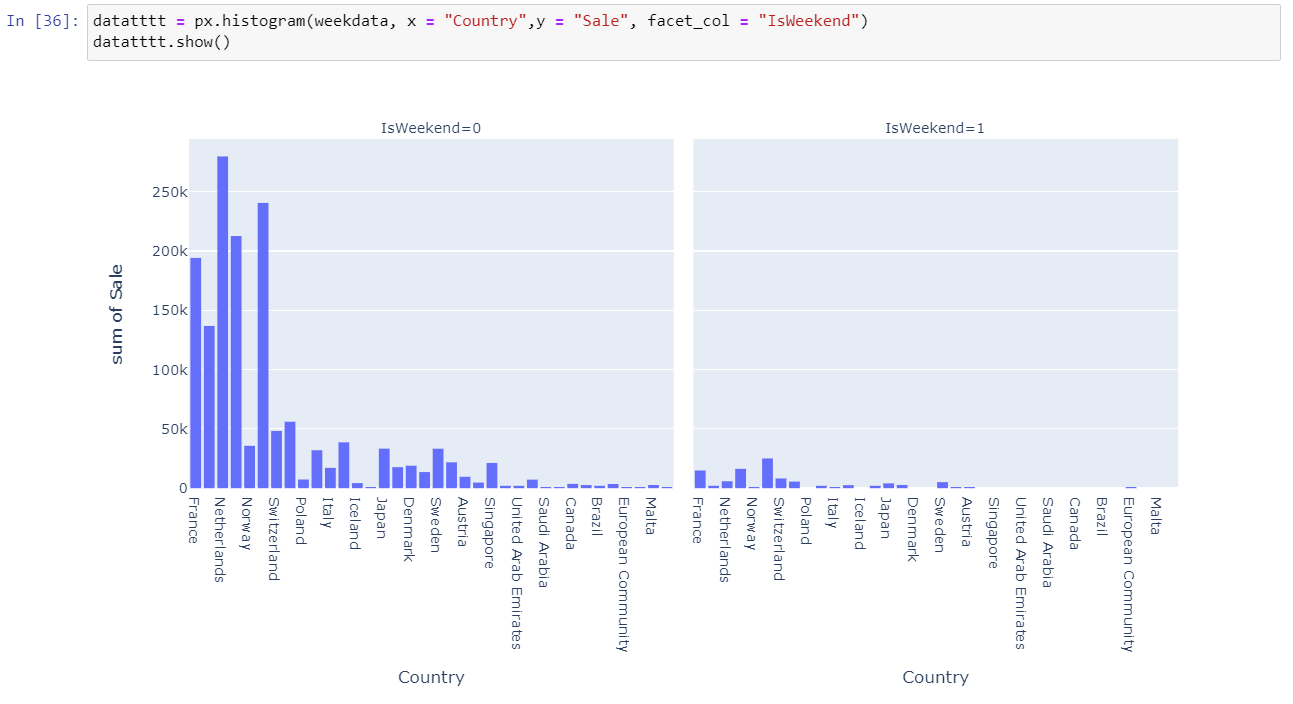




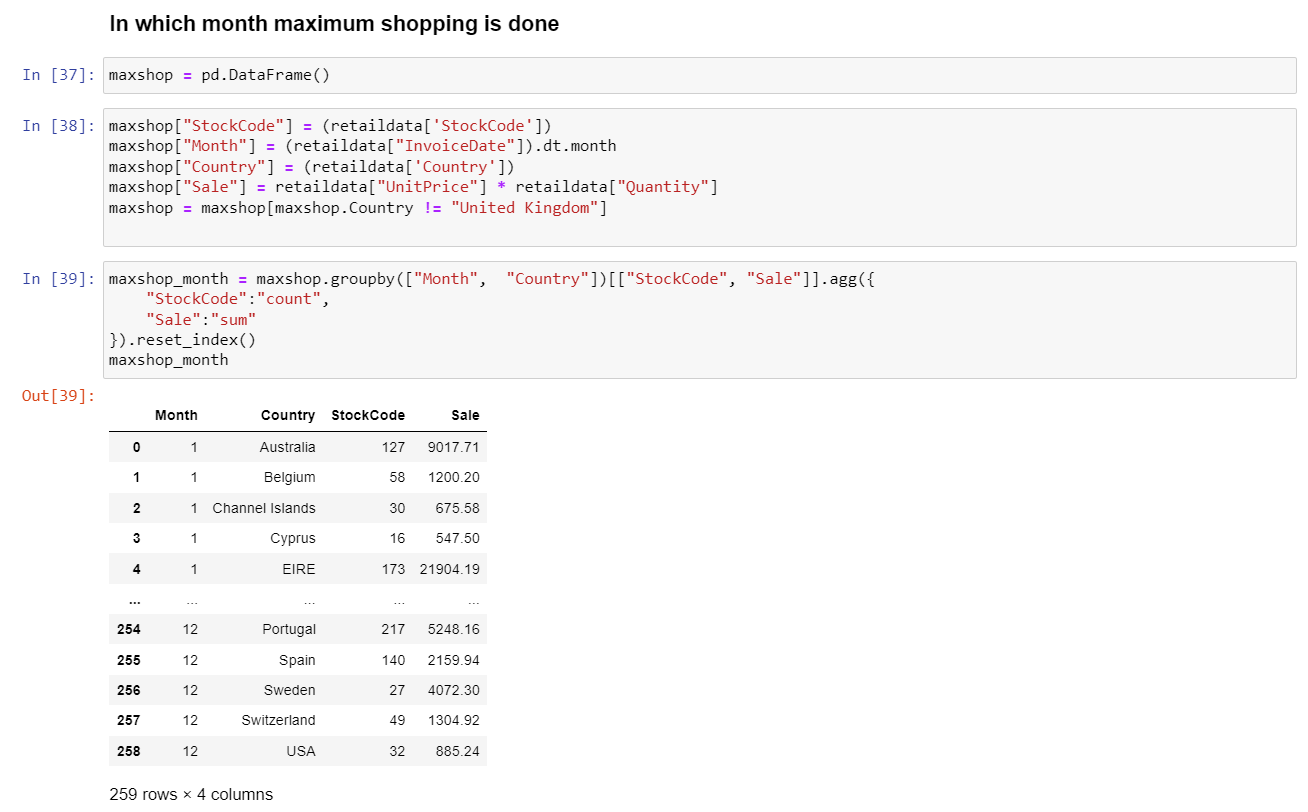


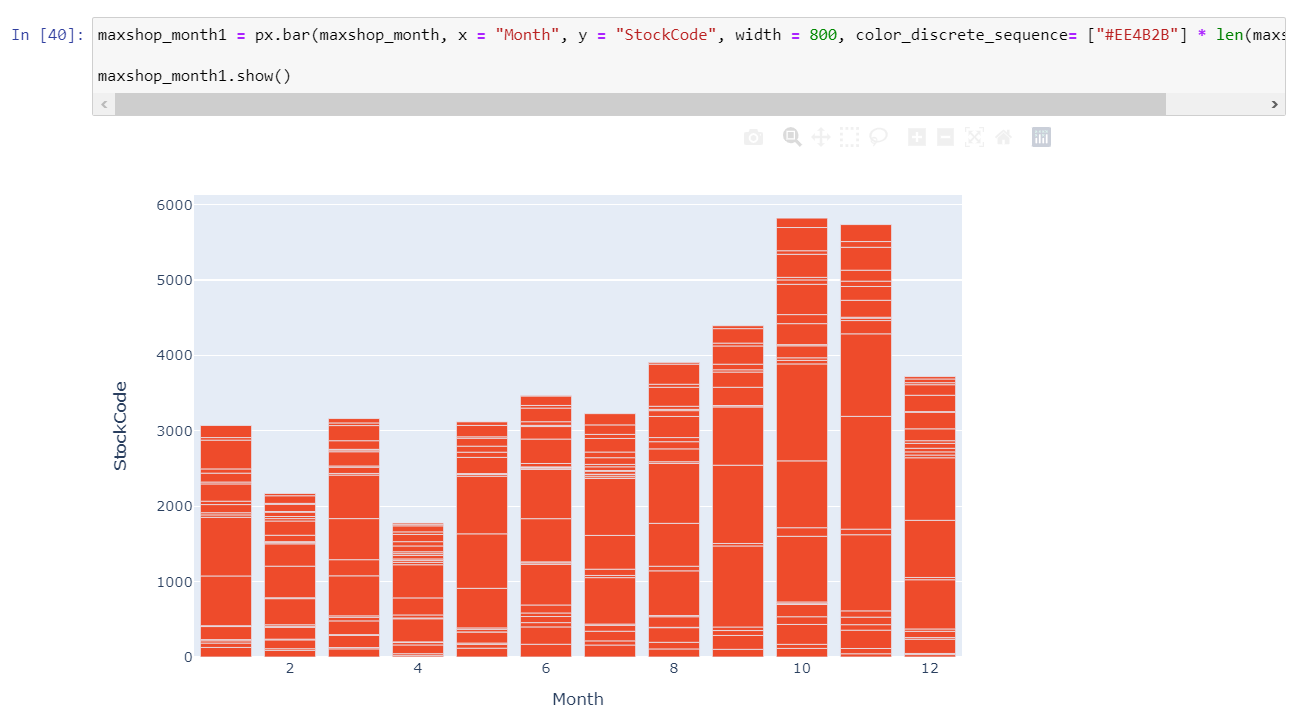


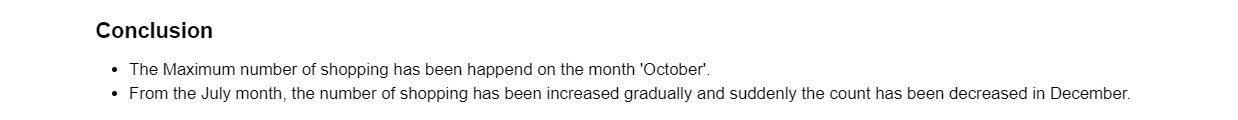




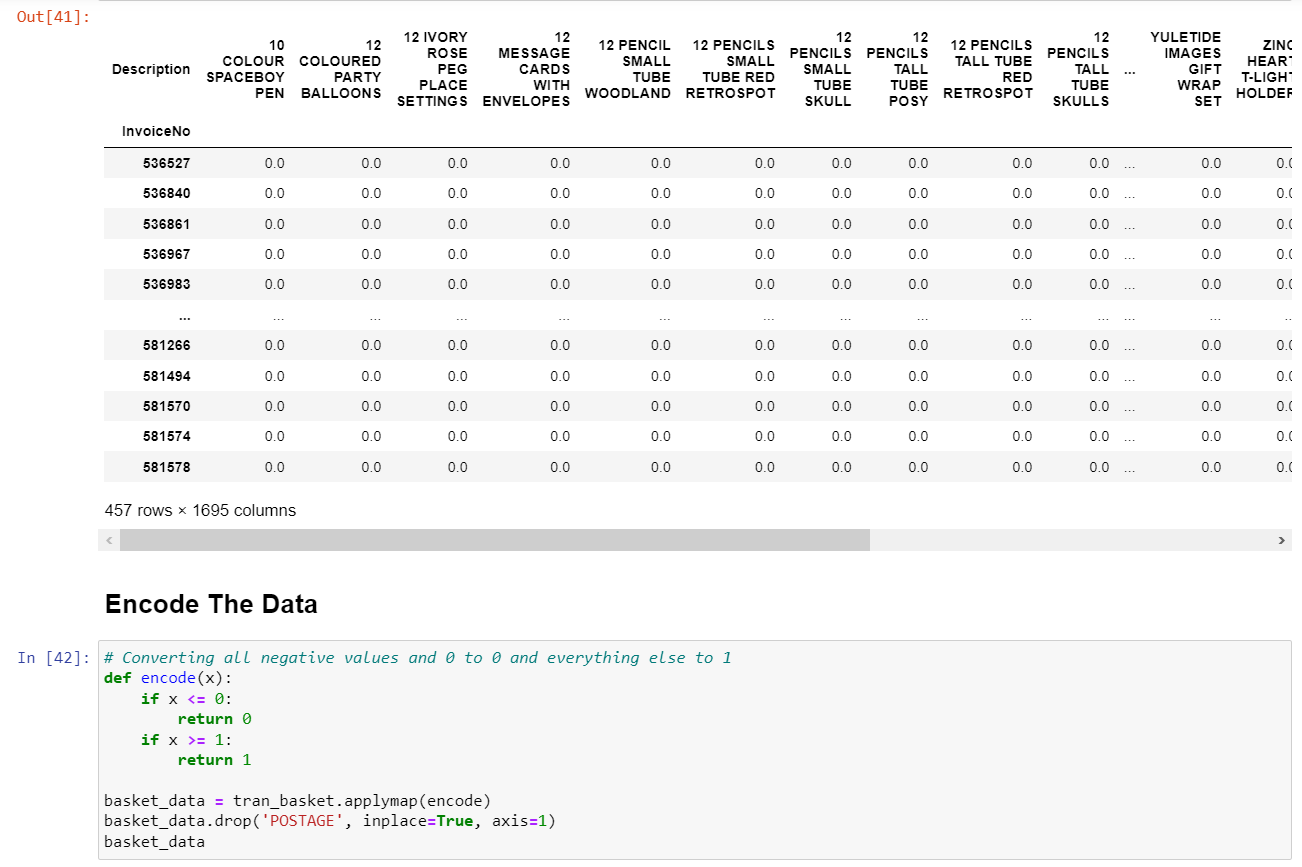




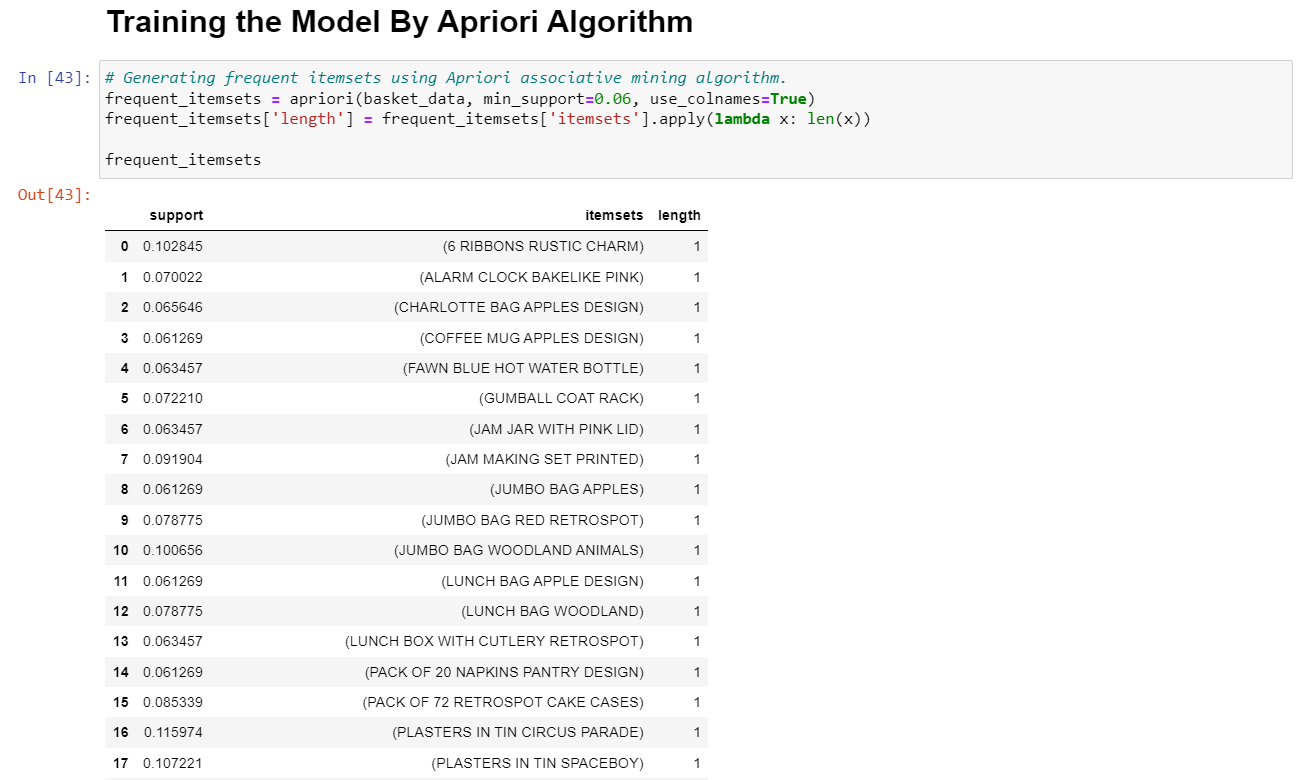




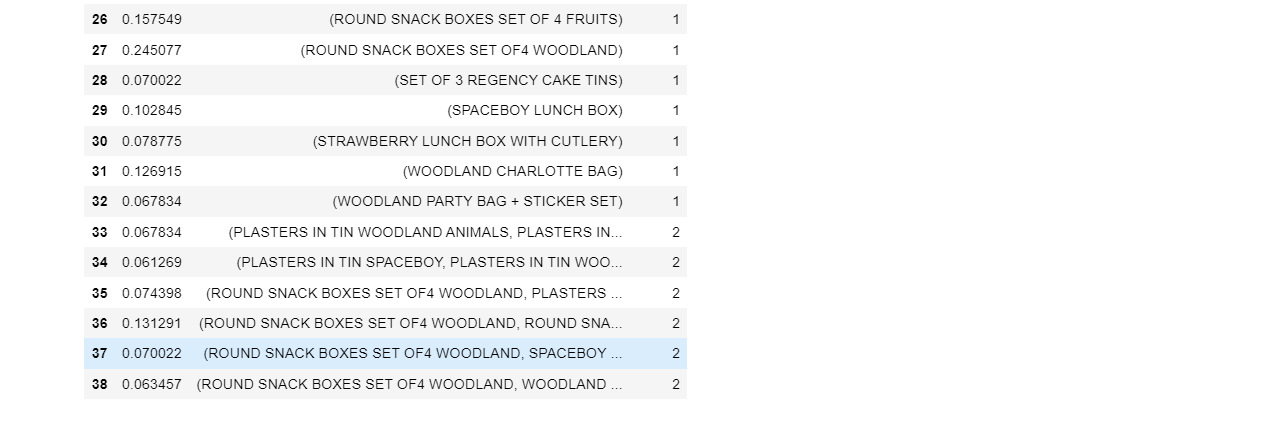


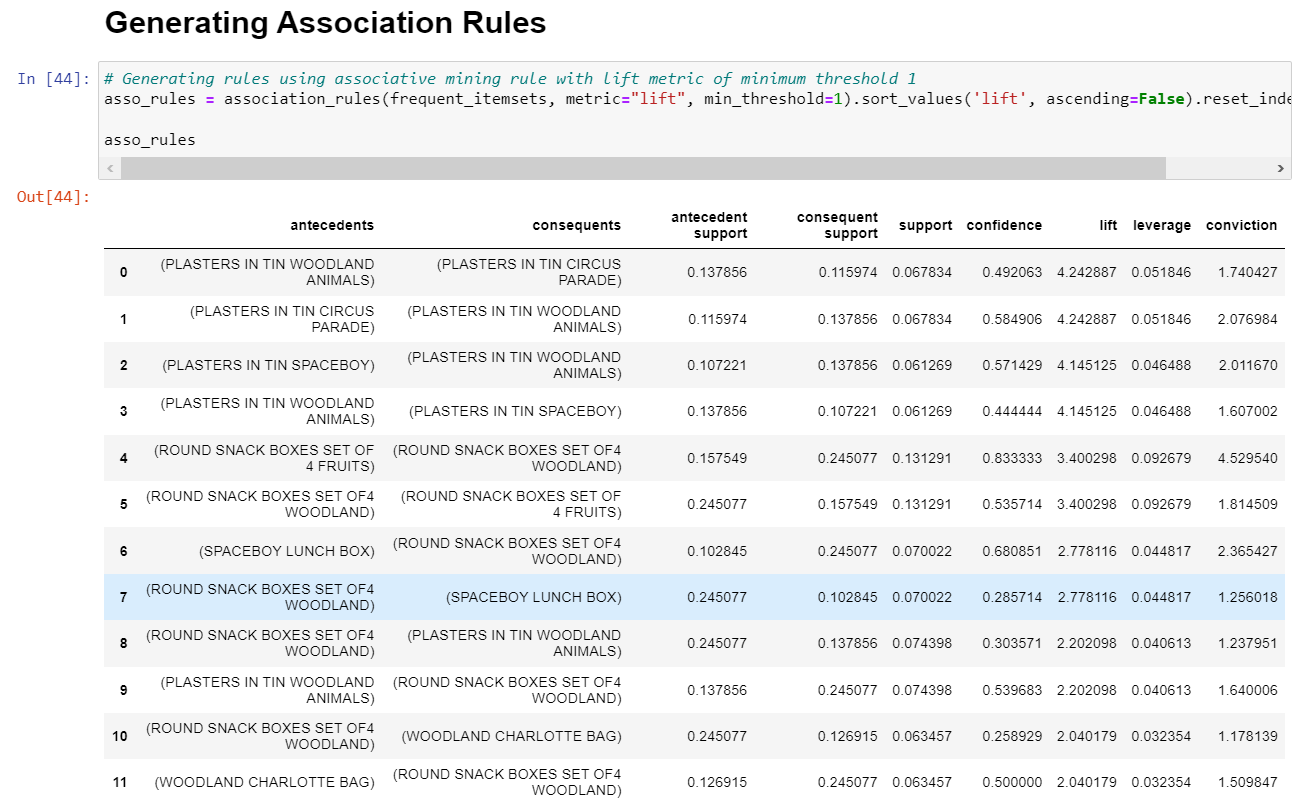


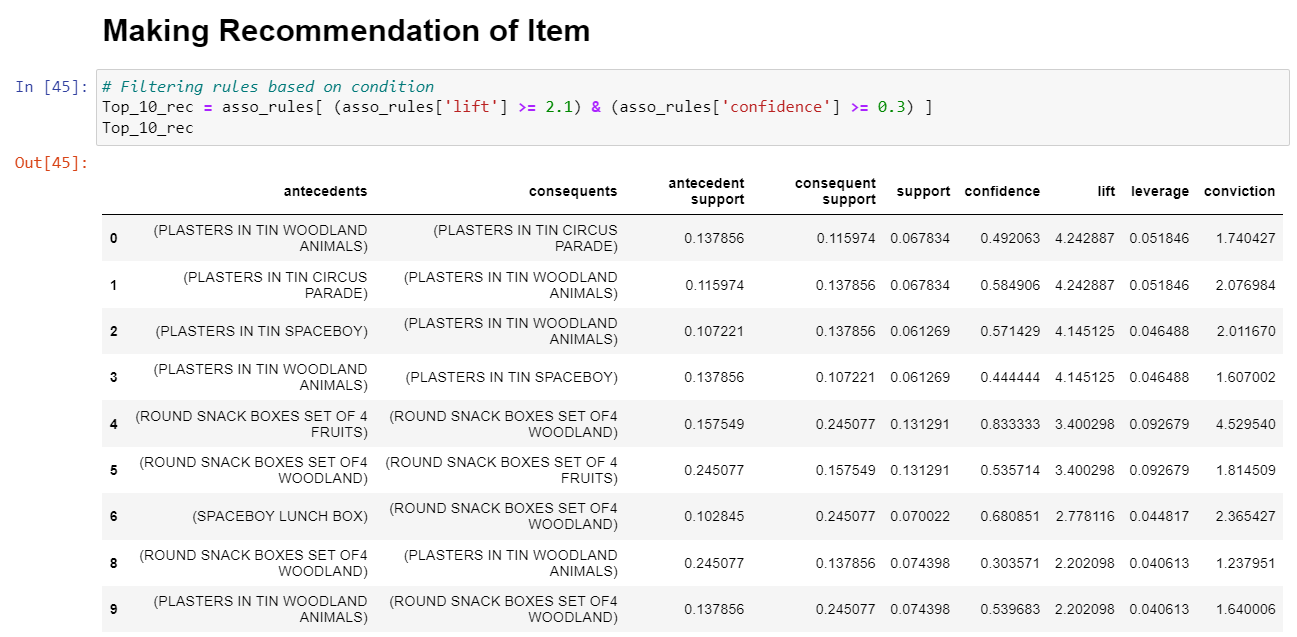


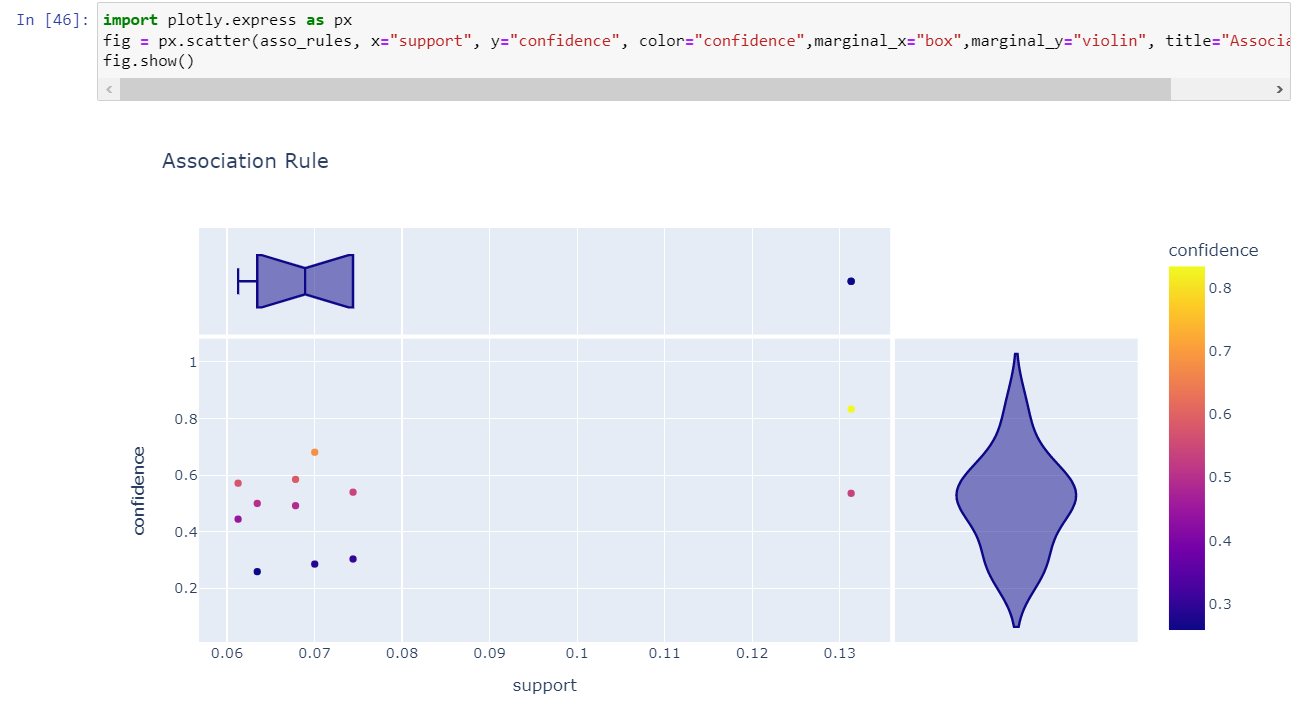


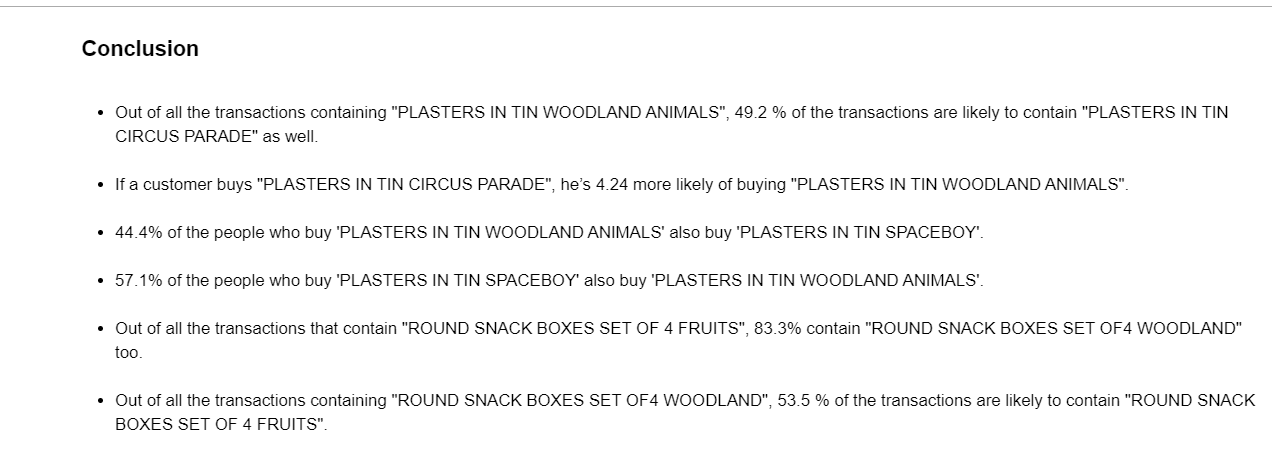


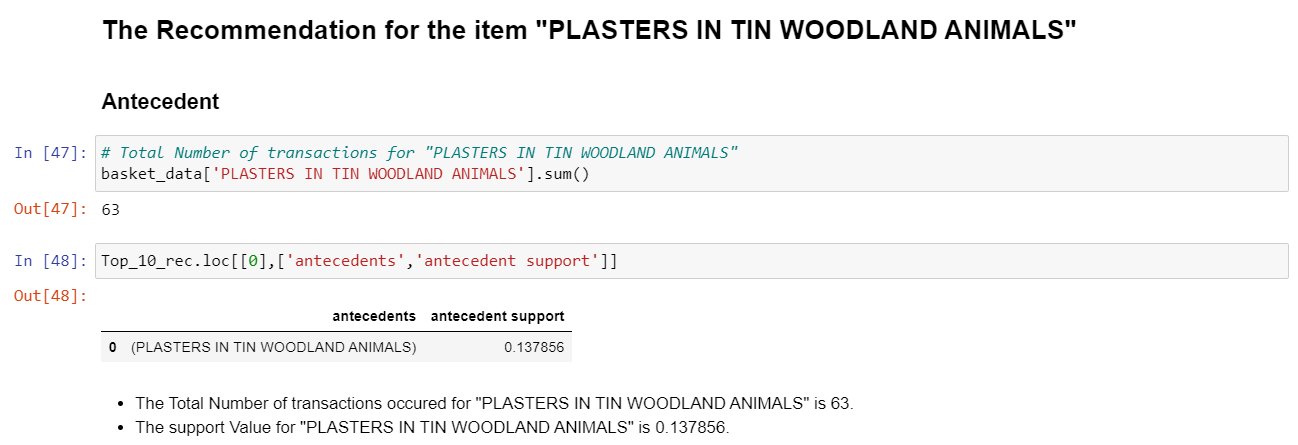
****

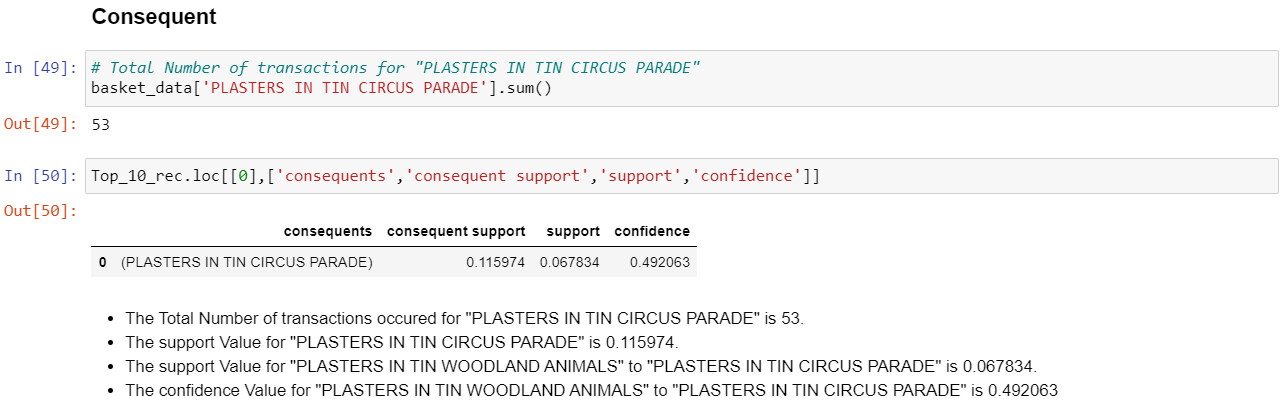
****

****

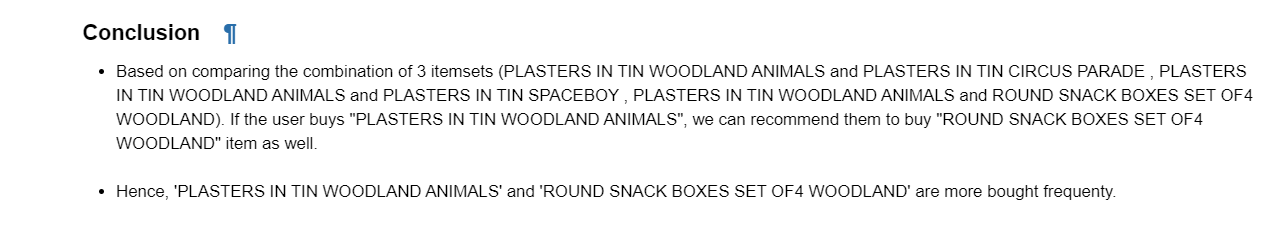
****

****

****

****

****

****

1. **RESULT**

**8.1 Conclusion**

Here we can conclude which are items that were most likely to be bought together. As a result, we may adapt our marketing and sales strategies to take advantage of this information. The items that were purchased together are listed below:

* Out of all the transactions including "plasters in tin woodland animals," 49.2% are likely to also include "plasters in tin circus parade."
* Customers who purchase "plasters in tin circus parade" are 4.24 times more likely to purchase "plasters in tin woodland animals."
* 57.1% of customers who purchase "plasters in tin space boy" also purchase "plasters in tin woodland animals."
* 83.3% of all transactions feature "round snack boxes set of 4 fruits" as well as "round snack boxes set of 4 woodland."
* 53.5% of all transactions including "round snack boxes set of 4 woodland" are likely to also contain "round snack boxes set of 4 fruits."

**8.2 Future Scope**

I aim to pursue the following work in the future:

* To determine how other factors affects the sales between the products.
* To determine how an item influence the customers to purchase another item.

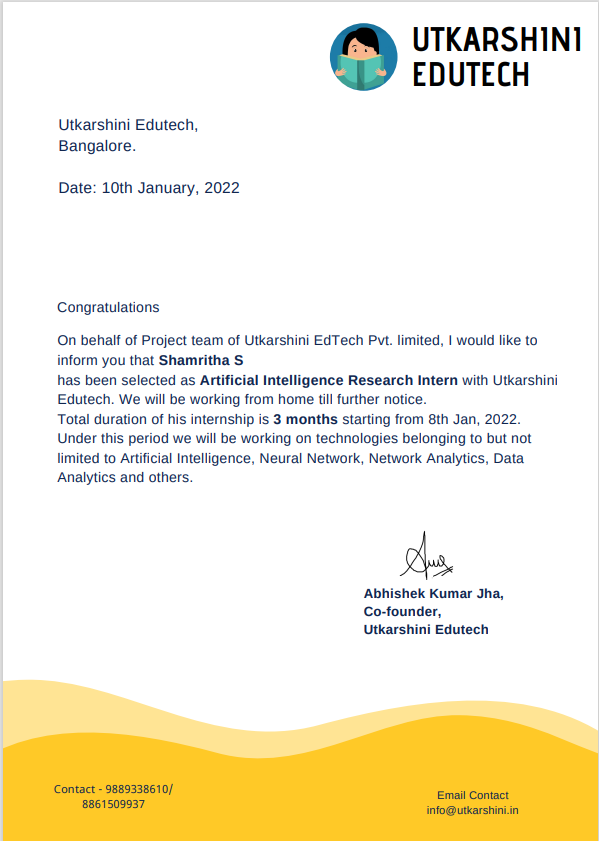
**9. REFERENCE**

* <https://sci-hub.hkvisa.net/10.1109/kamw.2008.4810675>
* https://sci-hub.hkvisa.net/10.1109/DMO.2012.6329808
* https://www.youtube.com/watch?v=txNvZ3Zndak
* https://sci-hub.hkvisa.net/10.1109/69.846291
* <https://link.springer.com/article/10.1057/dbm.2012.17>
* <https://sci-hub.hkvisa.net/10.1109/ICCTIDE.2016.7725377>
* <https://sci-hub.hkvisa.net/10.1109/DMO.2012.6329808>
* <https://sci-hub.hkvisa.net/10.1109/ICRTIT.2016.7569583>
* <https://sci-hub.hkvisa.net/10.1109/IC4.2015.7375725>
* <https://sci-hub.hkvisa.net/10.1109/INVENTIVE.2016.7830203>

**10. WORKLOG**

|  |  |  |
| --- | --- | --- |
| **Day** | **Date** | **Task Done** |
| Day 1 | 10/01/2022 | External: Referring the concept of Data Science creating the problem statement for the project. |
| Day 2 | 11/01/2022 | External: Collecting the data for the project. |
| Day 3 | 12/01/2022 | External: Working on data cleaning and data analysis part |
| Day 4 | 15/01/2022 | External: Working on data Preparation and data visualization part. |
| Day 5 | 16/01/2022 | External: Started learning about Association Mining with Apriori Algorithm |
| Day 6 | 18/01/2022 | External: Encode the Data, Training the Model by Apriori Algorithm. |
| Day 7 | 20/01/2022 | External: Encode the Data, Training the Model by Apriori Algorithm. |
| Day 8 | 22/01/2022 | External: Training the Model by Apriori Algorithm and working on the association rule mining. |
| Day 9 | 25/01/2022 | External: Training the Model by Apriori Algorithm and working on the association rule mining. |
| Day 10 | 28/01/2022 | External: Training the Model by Apriori Algorithm and working on the association rule mining. |
| Day 11 | 30/02/2022 | External: Making Pattern evaluation and data presentation. |
| Day 12 | 8/02/2022 | External: Attended external review meeting |

**11. OFFER LETTER**

****