Exploratory Data Analysis on Online Retail Transaction Data

2 Question 02

Introduction

The dataset represents the transactions of a UK based Online retail business over two years between 01/12/2009 to 09/12/2011. The business has its customers for various types of unique gift items worldwide and the dataset includes nearly one million customers. The data was provided in two separated excel sheets. The dataset has four different types of quantitative data such as 'Quantity', 'InvoiceDate', 'UnitPrice' and 'CustomerID', and four qualitative data namely as 'InvoiceNo', 'StockCode', 'Description', and 'Country'. The objective of this business analytics report is to investigate the relationship between these variables.

Data preparation

The data for the analysis is in two separate excel sheets so the separate sheets were first combined before further processing the data by concatenating. Python modules used for the data analysis include openpyxl, pandas, matplotlib and seaborn mainly.

The data set was then briefly observed to know about its content and nature. The columns and the data type of each column were identified by using the info() method on the loaded complete data set which was referred to as 'sheets combined'.

```
# Evaluate the quality of data
sheets_combined.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1067371 entries, 0 to 1067370
Data columns (total 8 columns):
 #   Column       Non-Null Count    Dtype
---  ------       --------------    -----
 0   Invoice      1067371 non-null  object
 1   StockCode    1067371 non-null  object
 2   Description  1062989 non-null  object
 3   Quantity     1067371 non-null  int64
 4   InvoiceDate  1067371 non-null  datetime64[ns]
 5   Price        1067371 non-null  float64
 6   Customer ID  824364 non-null   float64
 7   Country      1067371 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 65.1+ MB
```

Figure 1: Columns and Their Data Type

Next, the uniqueness of the columns was observed for each column to understand the existence of repeated values.

The data set contains improper data that has to be eliminated by using data cleaning techniques to ensure better results. One of the key techniques that was used included listing the null value count for each column.

```
# Identify null values
sheets_combined.isnull().sum()
```

```
Invoice             0
StockCode           0
Description      4382
Quantity            0
InvoiceDate         0
Price               0
Customer ID    243007
Country             0
dtype: int64
```

Figure 2: Null Values

From this information it was observed that the 'Description' and the 'Customer ID' columns had null values that has to account for during the analysis, for this study null values found in the 'Description' were replaced with the text 'Unknown' and the data rows with missing 'Customer IDs' were ignored during the analysis. The data set was then filtered to drop the duplicated data rows as well to further make the data set more reliable. The data set contained negative values for the quantity, which is highly unlikely to be valid data, so they were also removed from the dataset.

```
# Check the negative values in the 'Quantity' and 'Price' columns
print('Number of negative values in quantity : ', (sheets_combined.Quantity < 0).sum())
print('Number of negative values in price : ', (sheets_combined.Price < 0).sum())
```

```
Number of negative values in quantity :  18390
Number of negative values in price :  0
```

```
# Remove the rows with negative values in the 'Quantity' and 'Price' columns
sheets_combined = sheets_combined[(sheets_combined['Quantity'] >= 0) & (sheets_combined['Price'] >= 0)]
sheets_combined
```

Figure 3: Negative Values

The data set was described to be having data between 01.12.2009 to 09.12.2011 but when the data was filtered exceptions were found outside this range.

```
# Data outside the period 01.12.2009 to 09.12.2011
(sheets_combined.InvoiceDate >= '2011-12-09').sum()
```

```
611
```

```
# Data outside the period 01.12.2009 to 09.12.2011
(sheets_combined.InvoiceDate <= '2009-12-01').sum()
```

```
0
```

```
# Filter data between 01.12.2009 to 09.12.2011
sheets_combined = sheets_combined[(sheets_combined['InvoiceDate'] <= '2011-12-09') & (sheets_combined['InvoiceDate'] >= '2009-12-
sheets_combined
```

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085.0	United Kingdom
1	489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
2	489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085.0	United Kingdom
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom
...	...	...	...	...	...	...	...	...

Figure 4: Data Outside the Time Frame

These exceptions were removed to make sure that the data was only within the range specified.

Descriptive and Diagnostic Analysis

First, the demographic distribution of customers of this online retail business was analyzed and visualized as follows.

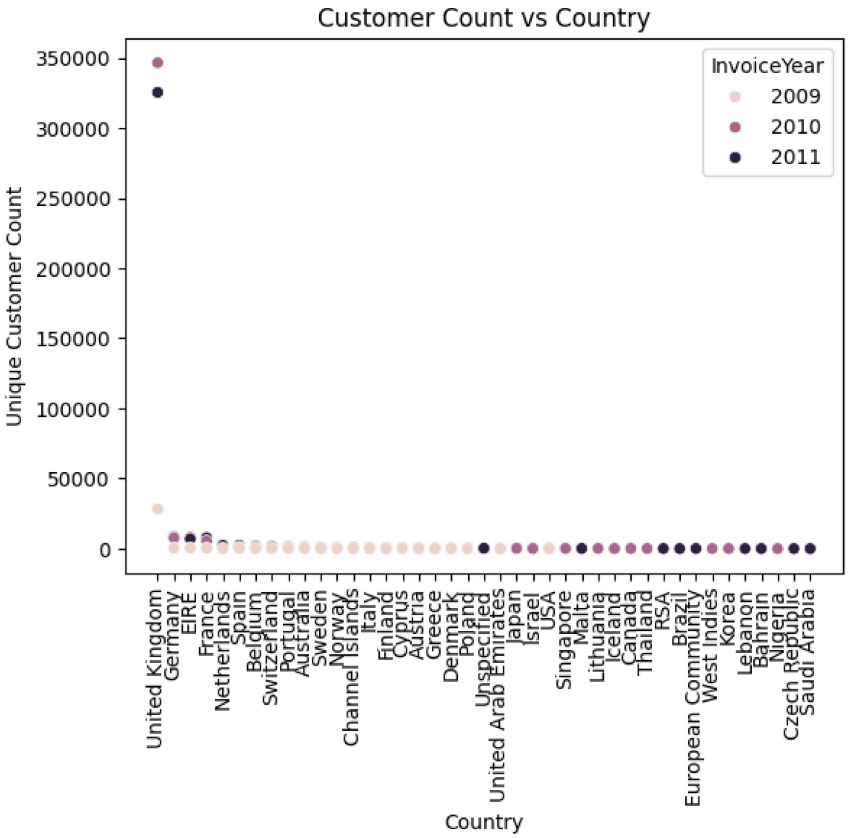

Figure 5: Customer Count vs Country

At a glance, It could identified that the country with the largest number of consumers from 2009 to 2011 is the United Kingdom. It has also shown a huge growth in the number of customers in the United Kingdom from 2010 and 2011. However, the distribution of consumers in other countries did not show significant change from 2009 to 2011.

Next, the revenue of the business by country was considered as follows.

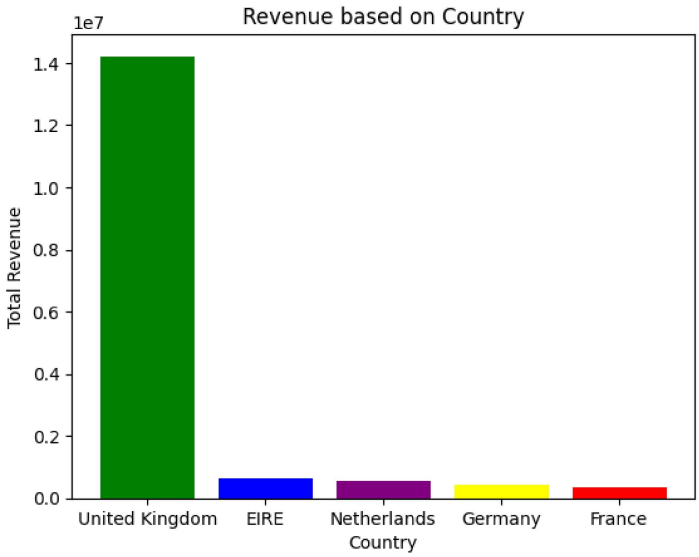

Figure 6: Revenue Based on Country

When the revenue by these countries were visualized it clearly showed that the United Kingdom accounted a significant proportion of the income compared to other four countries. The other countries, EIRE, Netherlands, Germany and France shows nearly equal amount of income.

The following figure shows the amount of goods that were bought from these countries.

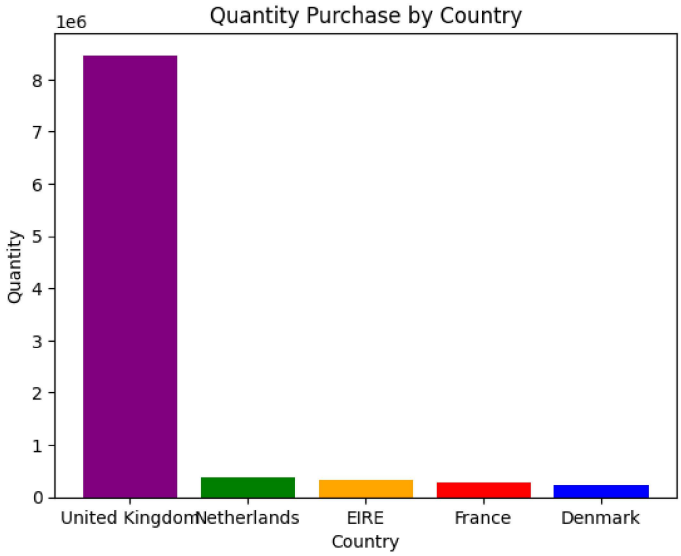

Figure 7: Quantity Purchase by Country

Plotting quantity against the countries, it could be observed that around 8 x 10^6 products were bought from the United Kingdom although Germany and France were the fourth and fifth largest contributors to the revenue of the business, France and Denmark were the fourth and fifth largest purchasers for the goods.

A scatter diagram was constructed to identify the customers who bought the highest quantity of products and their locations.

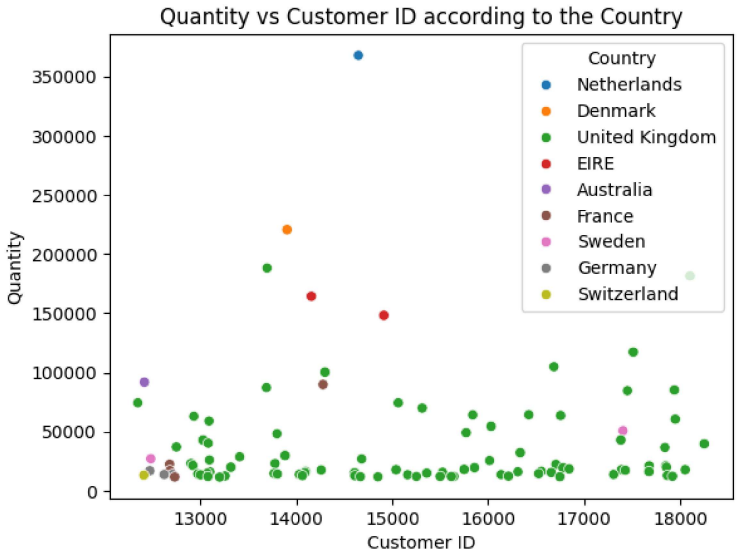

Figure 8: Quantity vs Customer ID According to the Country

Although the scatter plot shows that the UK has the highest number of consumers compared to other countries, Netherlands and Denmark respectively bought the most products.

The customers that gave the highest revenue for this online retail business were identified as follows and were visualized by the following pie chart.

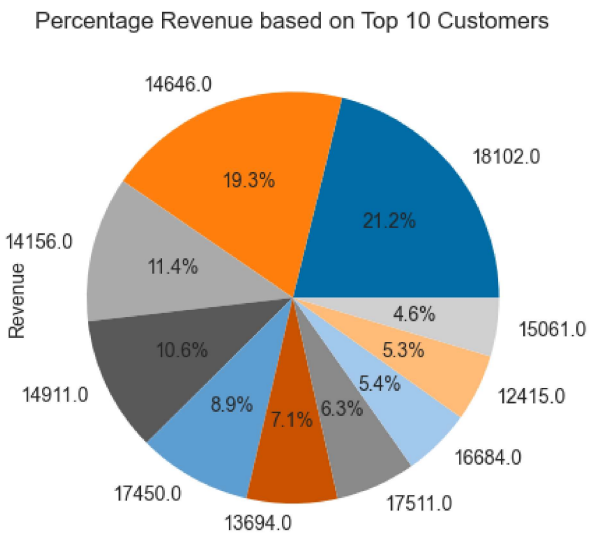

Figure 9: Percentage Revenue Based on Top 10 Customers

When the percentage of revenue based on the top 10 the customers were compared, the customer id number 18102.0 was the most profitable customer for the business with a contribution of 21% revenue and the customer id numbers 14646.0, 14156.0, 14911.0 and 17450.0 had also contributed to the business to generate more revenue respectively.

The customers who visit this business frequently were identified as follows.

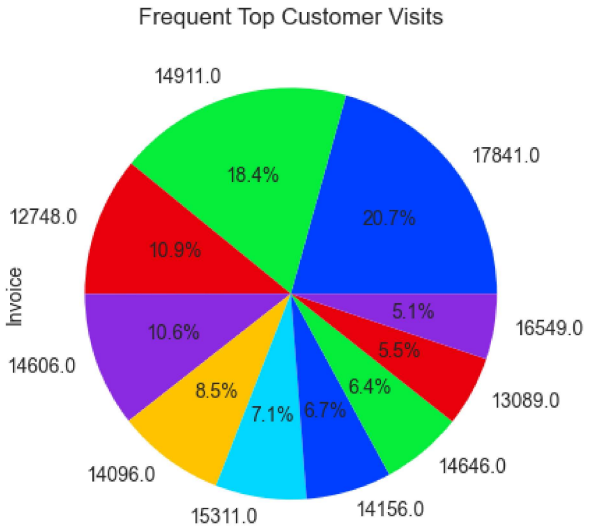

Figure 10: Frequent Top Customer Visits

According to the pie chart above the customer id numbers 17841.0, 14911.0, 12748.0, 14606.0 and 14096.0 were the customers who engage in this business and purchase goods more frequently. When the top 2 customers were compared based with the number of times, they purchased the products there was a slight difference, however other customers showed significant amount of difference when compared to these top 2 customers. With the analysis it was seen that not the customers who purchased most often made significance to drive the most revenue to this business.

The purchasing pattern of these top 5 customers over this period was as follows.

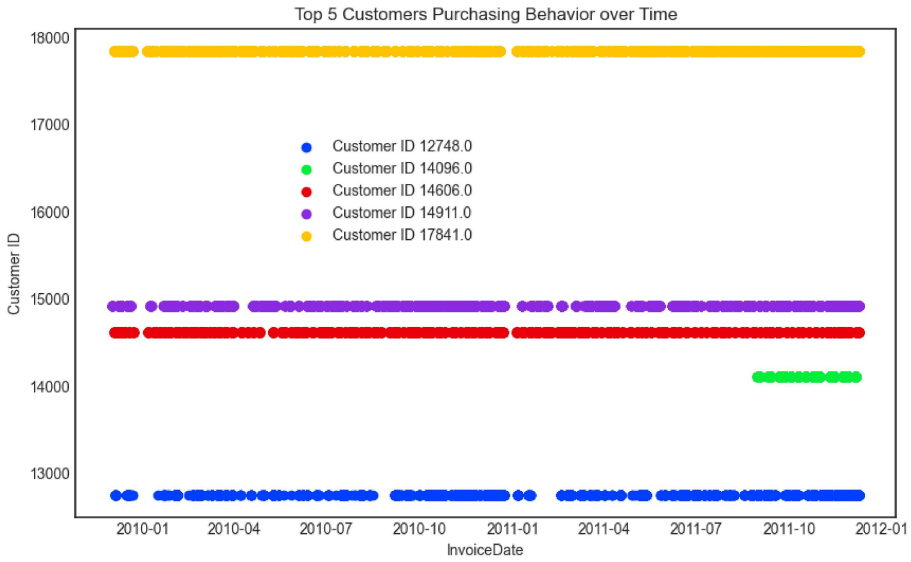

Figure 11: Top 5 Customers Purchasing Behavior Over Time

According to the scatter diagram it showed that all top 5 customers purchase from this business throughout the year except the customer id number 14606.0 as he was a new customer. For this business it can be recommended to have promotional campaigns, sale offerings, deals and discounts to encourage more customers.

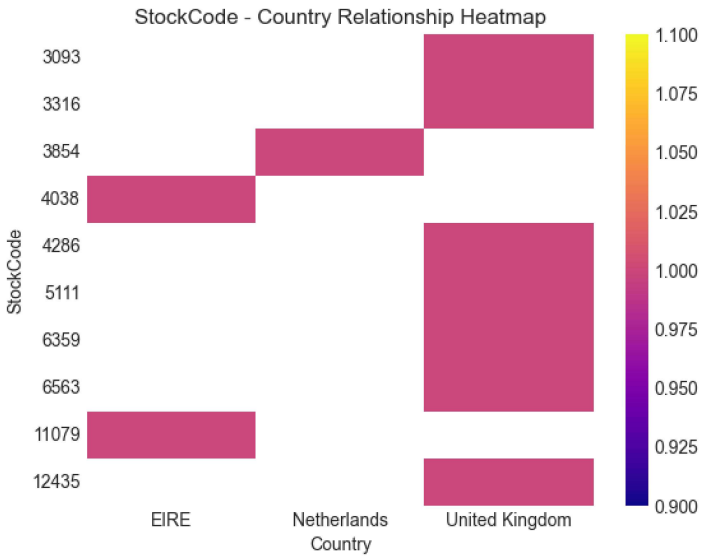

Figure 12: StockCode - Country Relationship Heatmap

A heatmap was used to find the relationship between stock code and country. According to the heatmap the United Kingdom have the highest relationship with stock code.

Predictive Analysis

Figure 13: Month Wise Revenue

When the revenue change over the past 2 years is considered, it was found that at the beginning of the year the revenue was shallow and in March it increased slightly and

again decreased and it fluctuated until August. From September it experienced a rapid surge to the end of the year. After that, when the stock code, quantity and customer id by month was considered it was found that they all followed the same pattern.

So, it could be stated that a specific reason behind this significant change could exist, and it can be predicted that seasonal changes have influence regarding this. However, more data is required to find the extract reason for this. This could provide more insights to promotional activities to spread the business among its customers. It could also be predicted that this same pattern will continue next year as well and allow the management of this firm to be prepared for a higher number of sales.

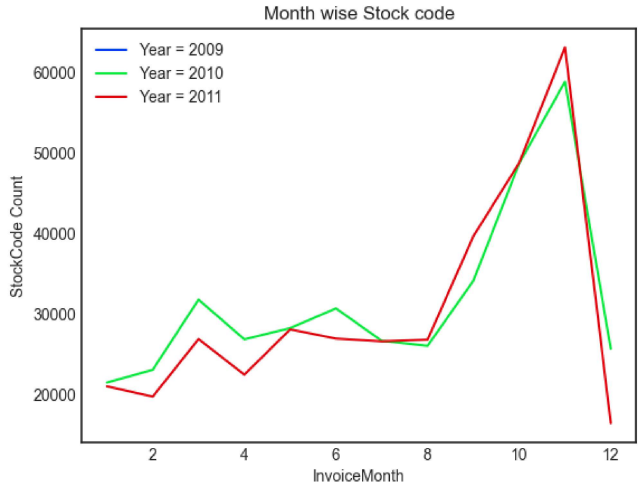

Figure 14: Month Wise Stock code

Figure 15: Sales Trends Over Time

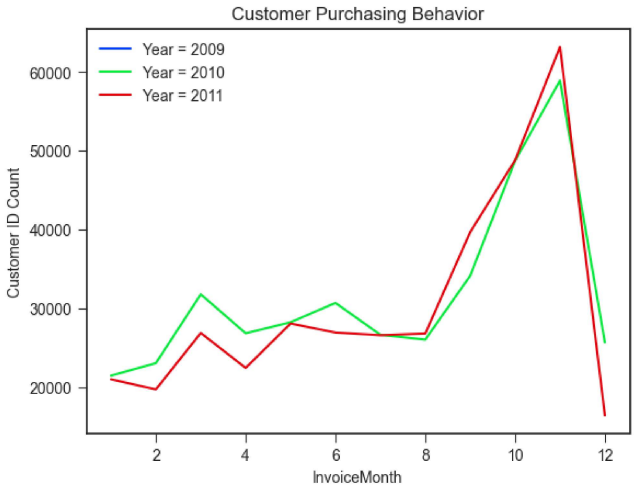

Figure 16: Customer Purchasing Behavior

When we compare the unit price with the quantity the bar graph below shows that the unit price of the item changes according to the number of quantities that were sold. The graph shows that the unit price was increased when the number of quantities decreased. So, if more units of the product could be sold next year, it would be possible to provide the product to the consumer at an even lower unit price.

Figure 17: Unit Price with Quantity

Two bar graphs were constructed to identify the products that were most popular and best-selling among customers in this online retail business. According to the graphs 'White Hanging Heart T-Light Holder' is the most popular and best-selling category, while 'Regency Cakestand 3 Tier' and 'Assorted Colour Bird Ornament' were the second and third respectively. So, if the same pattern continues for the coming years, it may be advisable to keep ready stock of these items to the required extent.

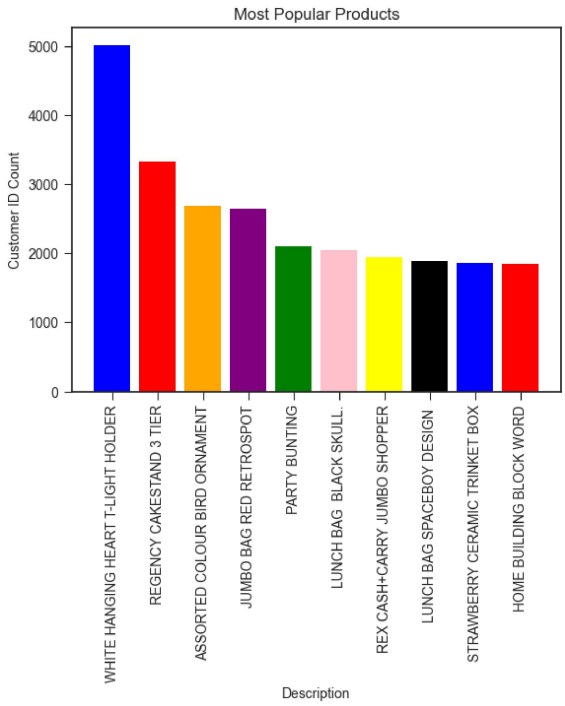

Figure 18: Most Popular Products

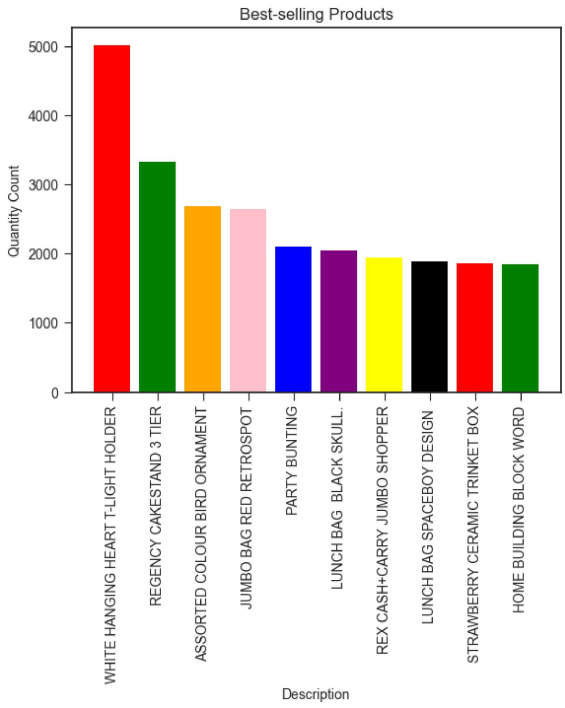

Figure 19: Best-selling Products

The following figure shows the correlation between several key features, there is a correlation of 0.7 between revenue and invoice date. At the same time, some features have a negative correlation and things like customer ID do not necessarily have a good correlation with anything.

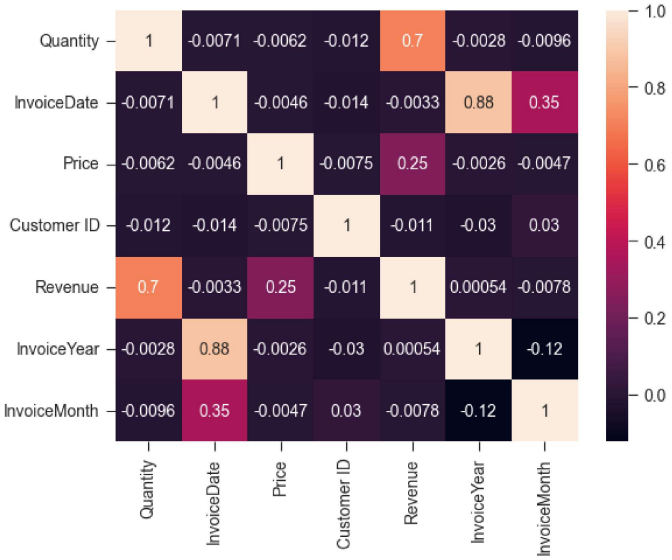

Figure 20: Correlations