Exploratory Data Analysis (EDA) Report

ONLINE RETAIL STORE

Table of Contents

1 Introduction

The dataset titled "Online Retail II" encompasses all the transactions that took place between 01/12/2009 and 09/12/2011 for an online retail business situated in the United Kingdom. The business is registered and operates solely online, without any physical store locations. The primary focus of the company is the sale of distinctive gift-ware suitable for various occasions. A significant portion of the company's clientele consists of wholesale buyers.

Exploratory Data Analysis (EDA) is an essential and fundamental stage within the data analysis workflow. Data analysis can aid in the identification of evident errors, as well as enhance comprehension of patterns within the data, discover outliers or unusual events, and uncover intriguing relationships among variables. The utilisation of a thorough EDA report facilitates the comprehension of the Online Retail II dataset, enabling the identification of patterns and the acquisition of insights prior to engaging in more sophisticated analyses or modelling techniques.

2 Data Exploration

2.1 Summary of the basic properties of the dataset.

Column	Description	Data Category	Data Type	Default Data Pattern
Invoice	Invoice number	Nominal	object	A 6-digit integral number is uniquely assigned to each transaction.
StockCode	Product (item) code	Nominal	object	A 5-digit integral number is uniquely assigned to each distinct product.
Description	Product (item) name	Nominal.	object	Product name
Quantity	The quantities of each product (item) per transaction.	Numeric	int64	Quantity in Numbers
InvoiceDate	Invoice date and time	Numeric	datetime64[ns]	The day and time when a transaction was generated.
Price	Unit price	Numeric	Float64	Product price per unit in sterling (Â£).
Customer ID	Customer number.	Nominal.	Float64	A 5-digit integral number is uniquely assigned to each customer.
Country	Country name.	Nominal	Object	The name of the country where a customer resides.

2.2 Summary of the observation of the dataset.

Column	No of entries	Unique values	Missing values	Negative values	Zero values	Outliers Values	Inconsistent coding
Invoice	1,067,371	53,628	0	N/A	N/A	N/A	Identified
StockCode	1,067,371	5,305	0	N/A	N/A	N/A	Identified
Description	1,062,989	5,698	4,382	0	0	N/A	Identified
Quantity	1,067,371	1,057	0	22,950	0	Identified	N/A
InvoiceDate	1,067,371	4,7635	0	0	0	N/A	Unidentified
Price	1,067,371	2,807	0	5	6,202	Identified	N/A
Customer ID	824,364	5,942	243,007	0	0	N/A	Unidentified
Country	1,067,371	41	0	0	0	N/A	Identified

3 Data preprocessing

3.1 Data cleaning

#	Steps	Description
1	Delete canceled/adjusted invoices	Dataset initially had 1,067,371 entries, Removed 19,500 of which were cancelled invoices and adjusted bad debt.
2	Eliminating conflicting stock codes	Removed 4,879 rows.
3	Removing "Unspecified" Country	Removed 752 rows.
4	Removing missing values in the Description	Removed 4,319 rows.
5	Finding negative quantities	Replaced 763 rows with positive values.
6	Zero-value quantity identification	No zeros
7	Removing quantity outliers	Removed 1,619 outlier rows
8	Removing price zeros	Found 1,622 rows and removed

9	Replacing negative price values	Zero-negative values
10	Removing price outliers	Removed 7,286 outlier rows
11	Detecting description coding errors	Few products identified, no action done.
12	Finding inconsistent Invoice Date coding	Unidentified. End of the data cleaning process 1,025,832 rows ready for analysis.

3.2 Data transformation

In order to enhance the depth of the data analysis, a new column, denoted as "Revenue," has been incorporated into the dataset. This column is calculated by multiplying the quantity of items sold by their respective unit prices. Inserted Year, Month, Day, Hour columns by transforming the InvoiceDate column.

3.3 Date reduction

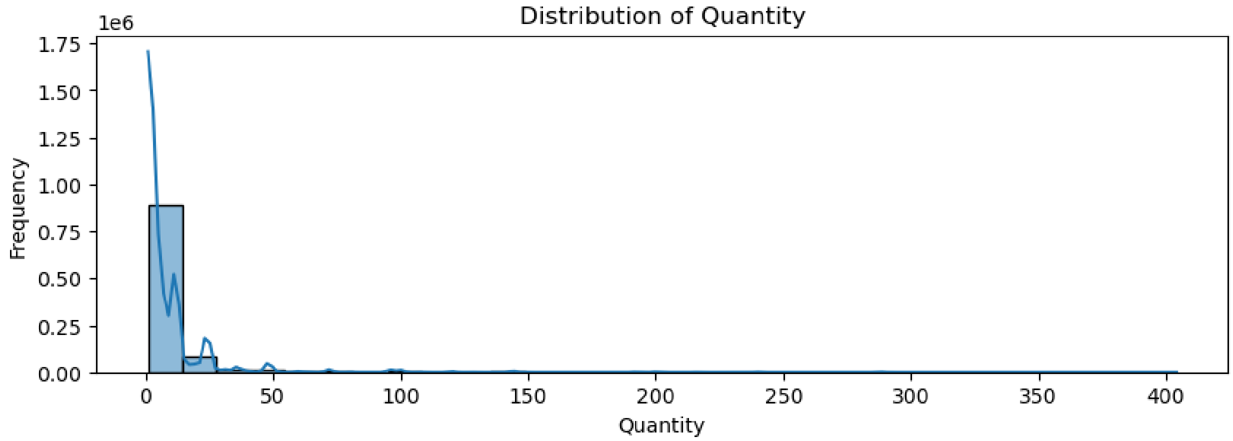
The fields "Invoice" and "StockCode" were excluded as they solely serve the purpose of product identification and lack significant content to yield valuable insights. The "InvoiceDate" columns were previously divided into four independent columns and excluded from the dataset for analysis. The "Customer ID" column contains 235,749 missing data points, which accounts for approximately 25% of the total data rows. Therefore, this column is likewise excluded from the dataset.

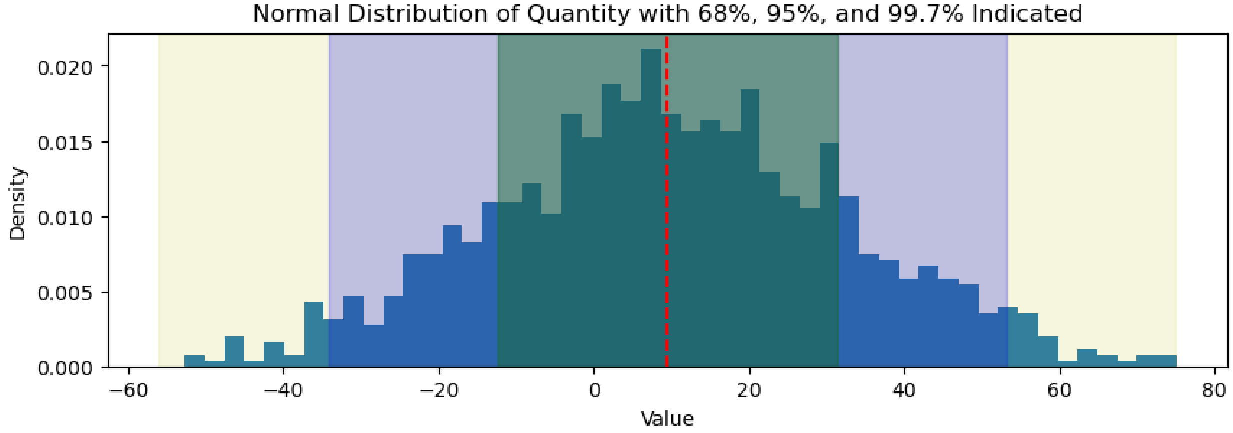
4 Data Analysis and Visualisation

4.1 Univariate analysis

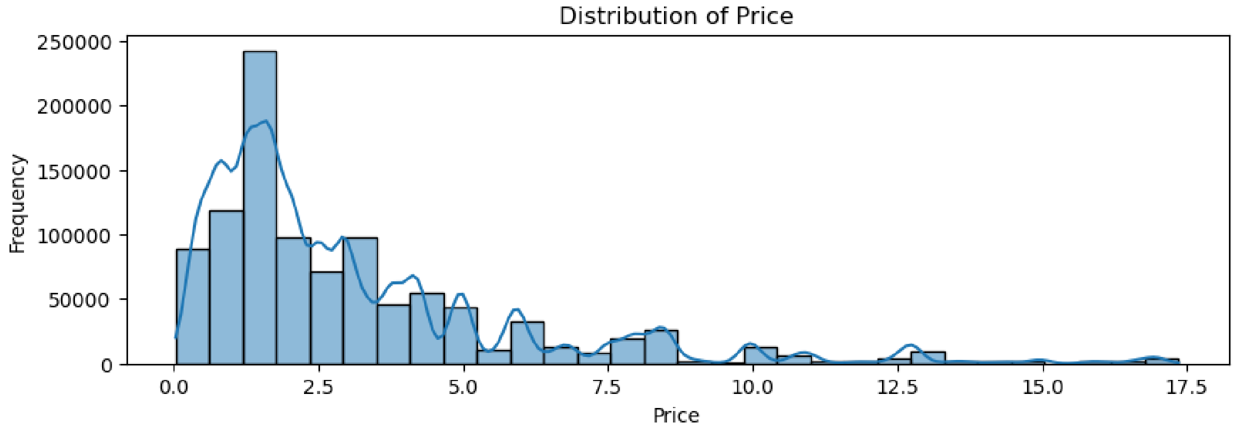
	Count	Unique values	mean	std	min	25%	50%	75%	max	IQR
Quantity	1025832	291	9.5	21.8	1	1	3	11	404	10
Price	1025832	546	3.14	2.94	0.04	1.25	2.10	4.13	3.05	2.88
Revenue	1025832	5054	17.4	45.01	0.06	3.90	9.95	17	3941.99	13.1

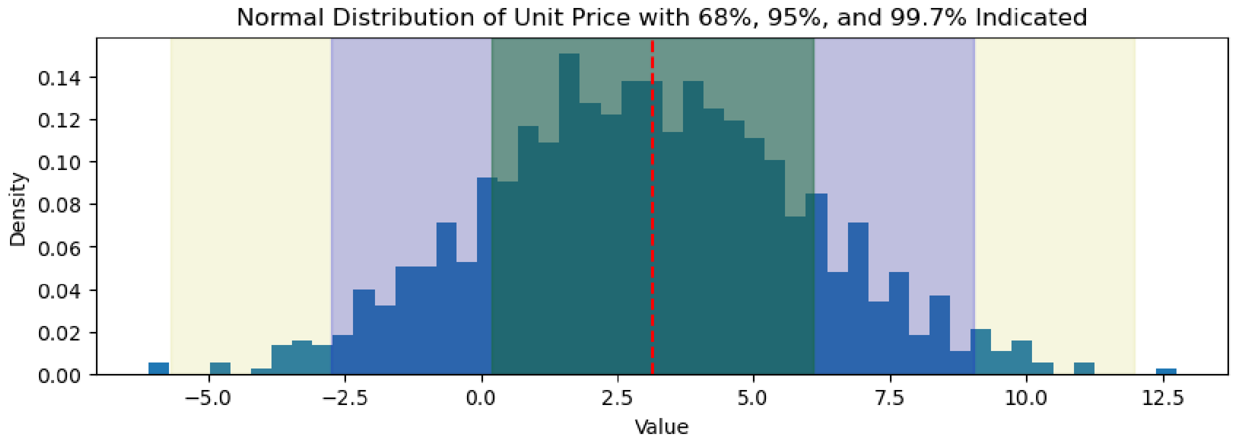
4.1.1 Quantity



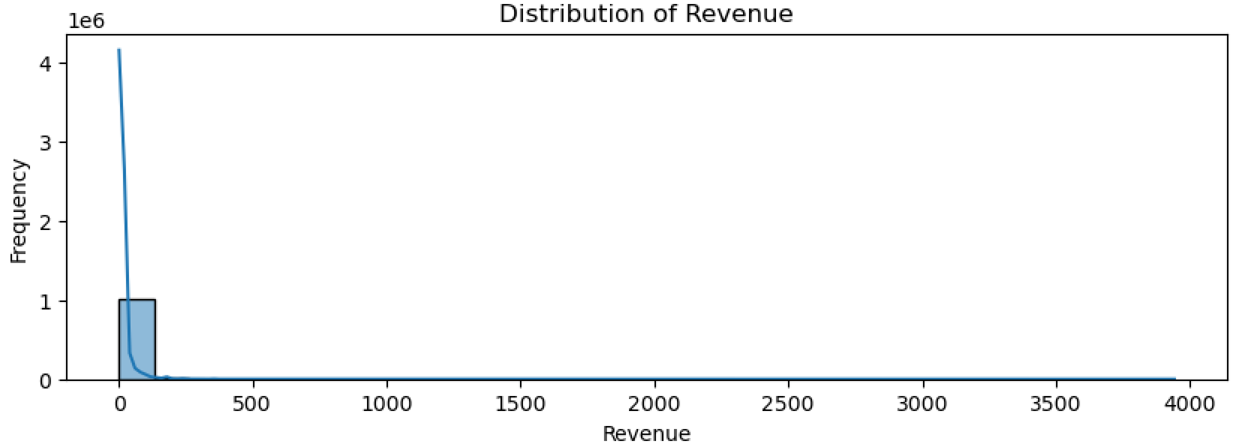


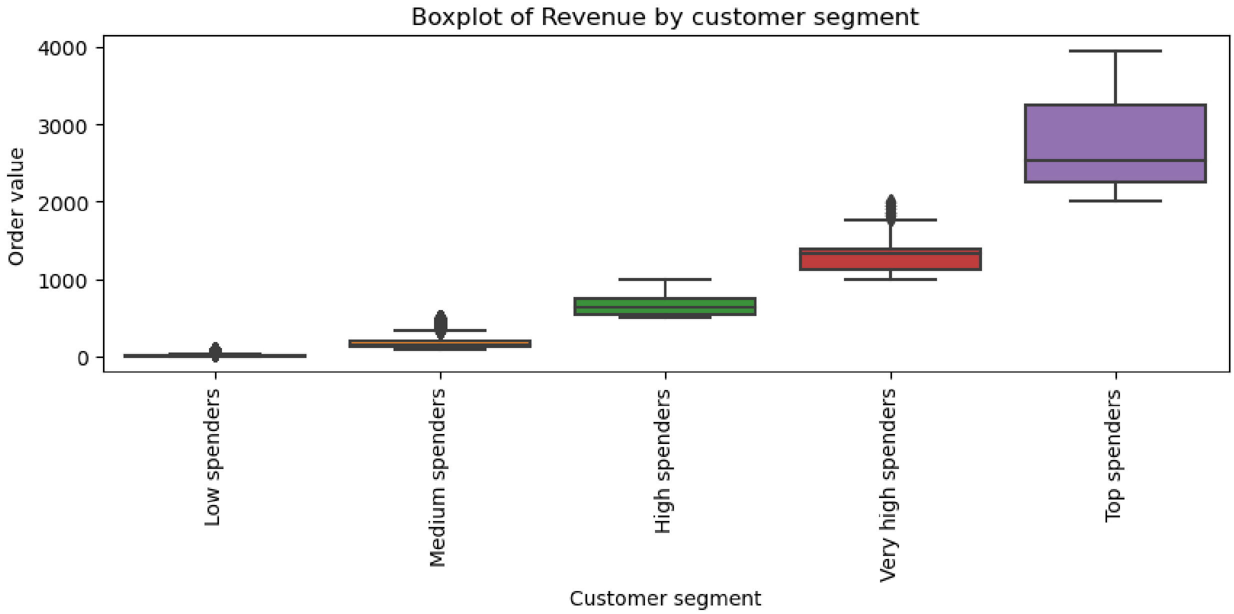
4.1.2 Unit Price

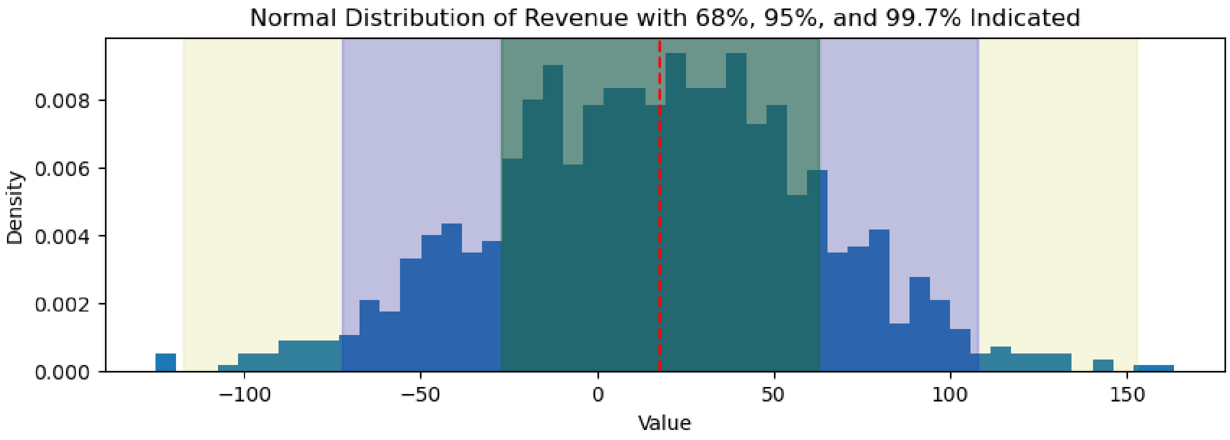



Normal Distribution of Unit Price with 68%, 95%, and 99.7% Indicated
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
Value

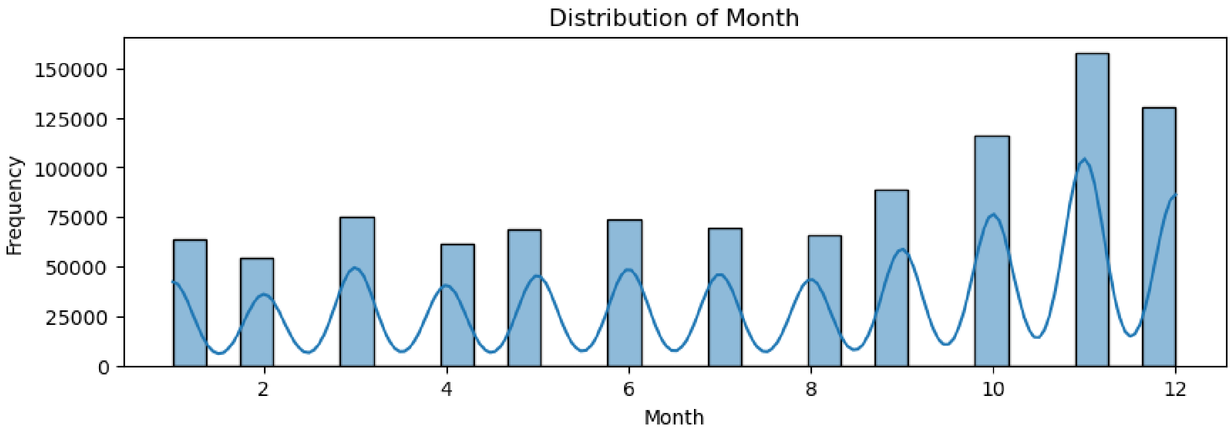
4.1.3 Revenue


Distribution of Revenue
1e6
Frequency
0
1
2
3
4
0
500
1000
1500
2000
2500
3000
3500
4000
Revenue

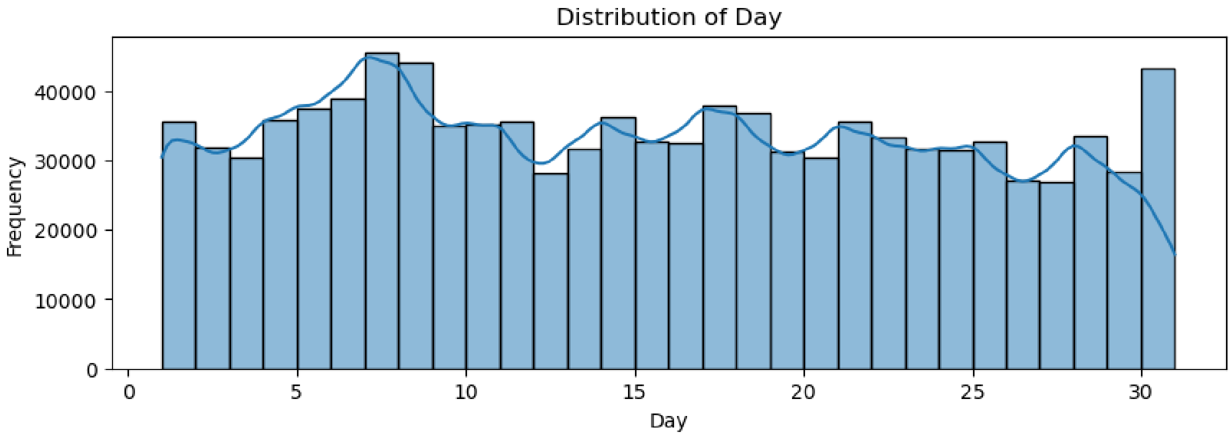

Boxplot of Revenue by customer segment
Order value
0
1000
2000
3000
4000
Low spenders
Medium spenders
High spenders
Very high spenders
Top spenders
Customer segment



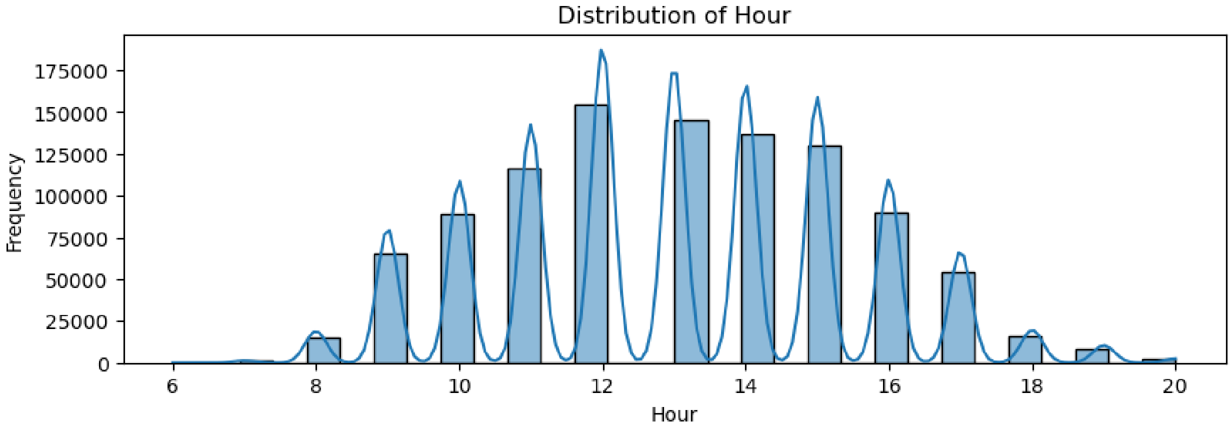
4.1.4 Month



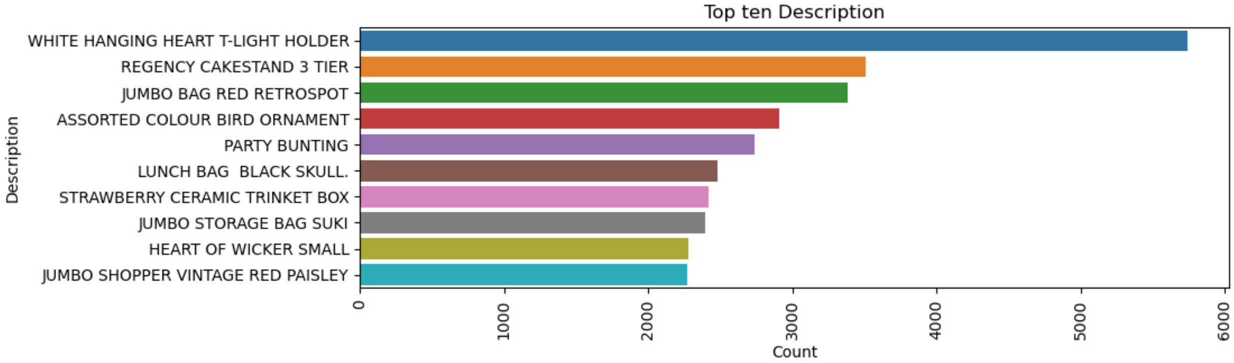
4.1.5 Day



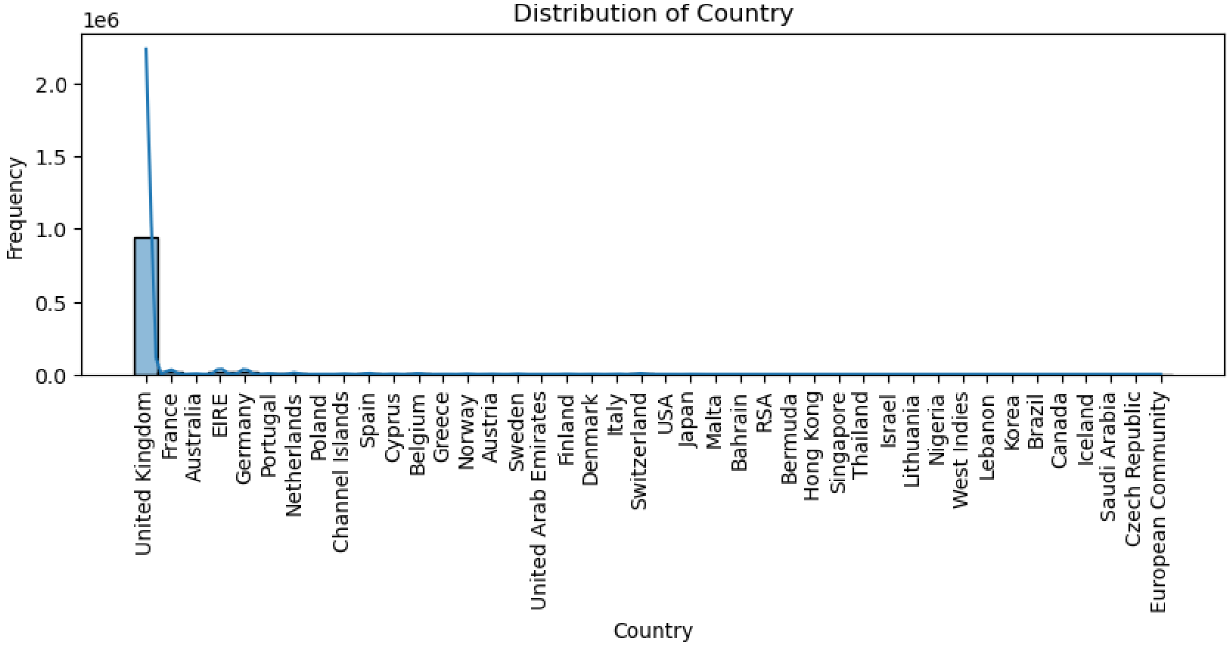
4.1.6 Hour

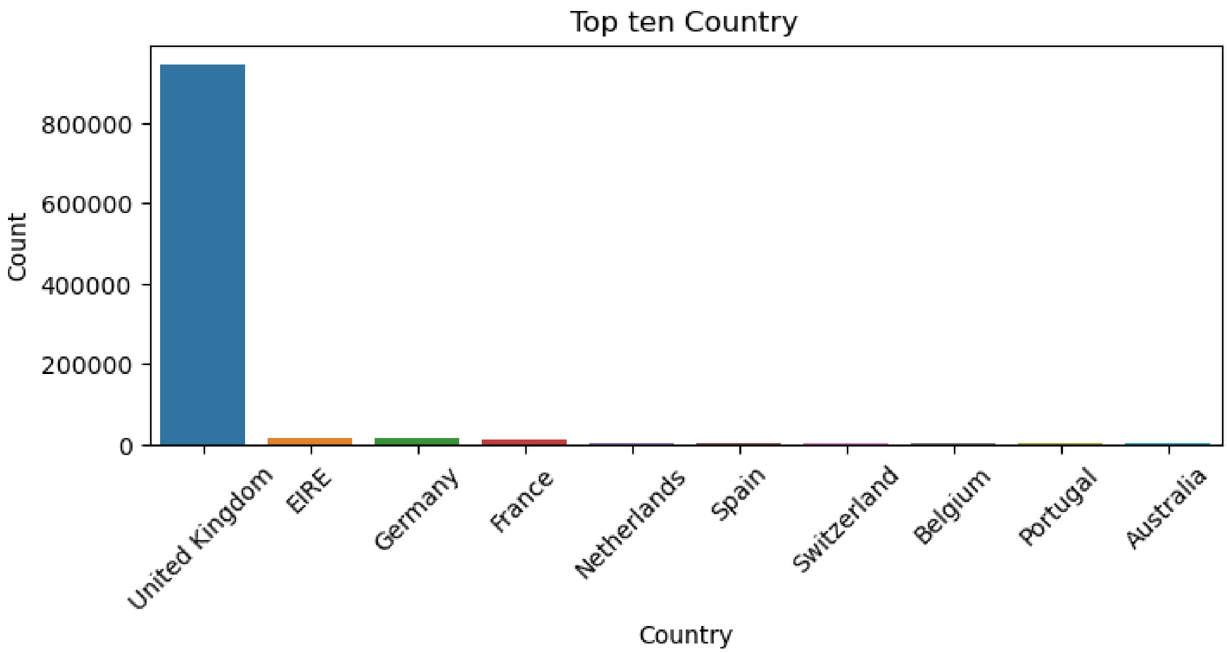


4.1.7 Product Description



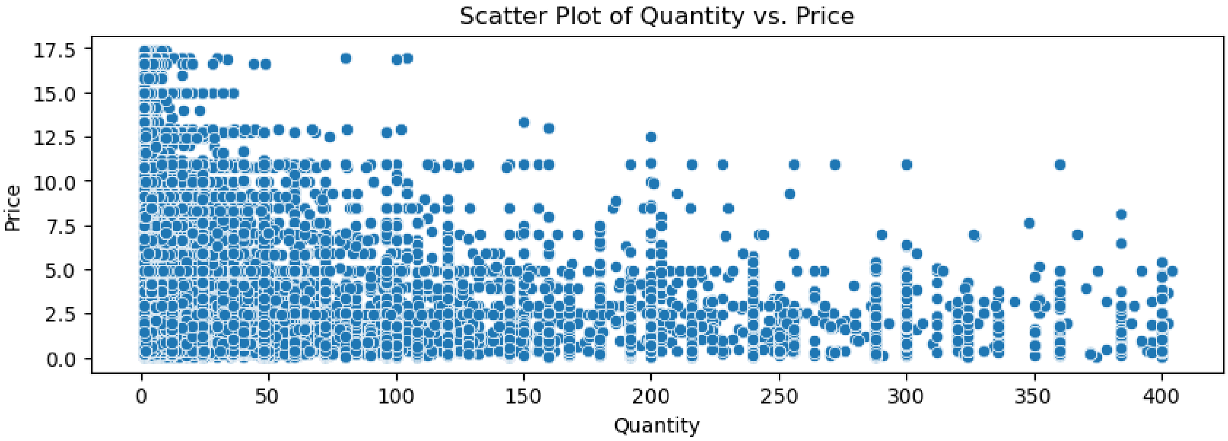
4.1.8 Country

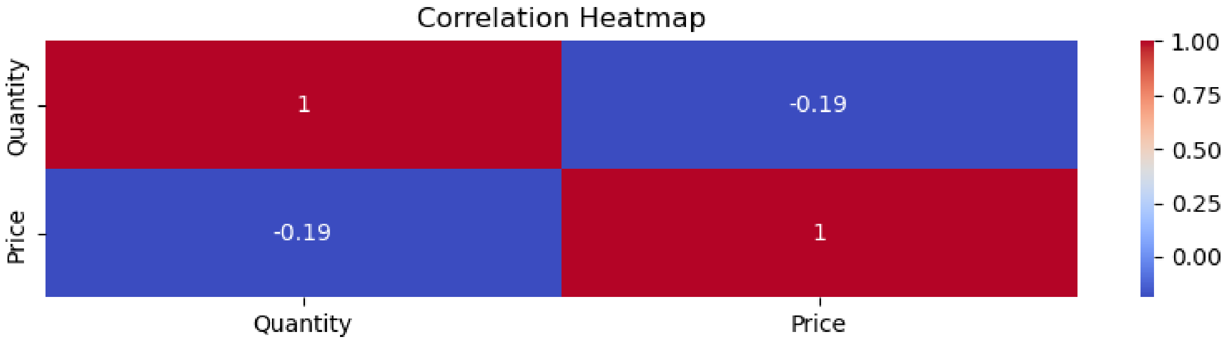


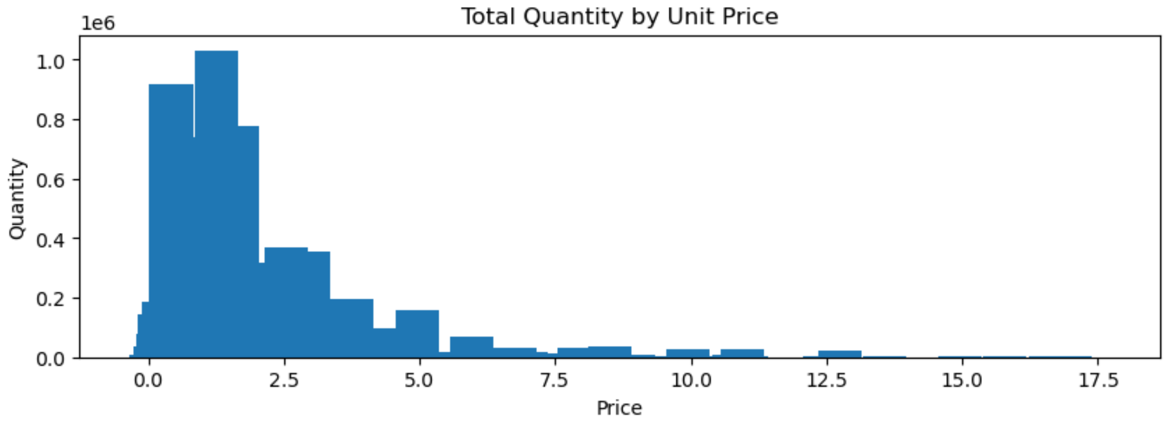


4.2 Bivariate Analysis for numeric vs numeric

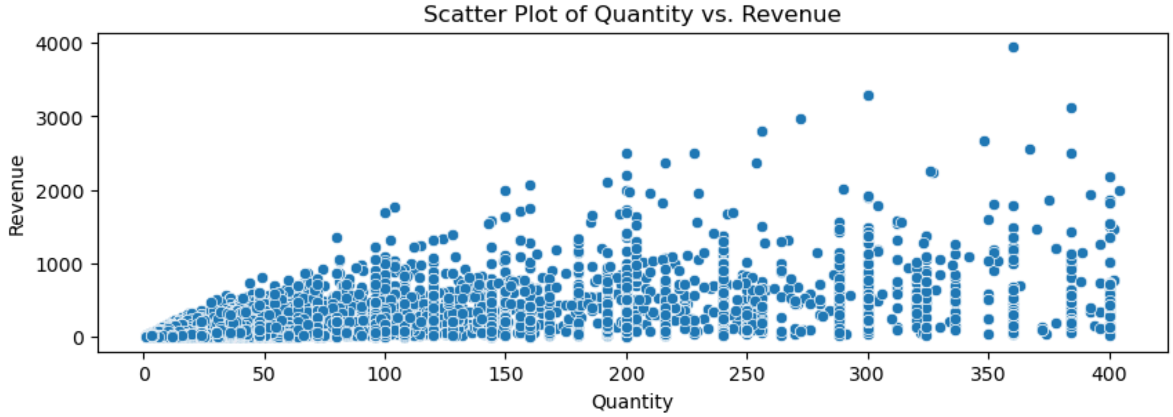
4.2.1 Quantity Vs Price

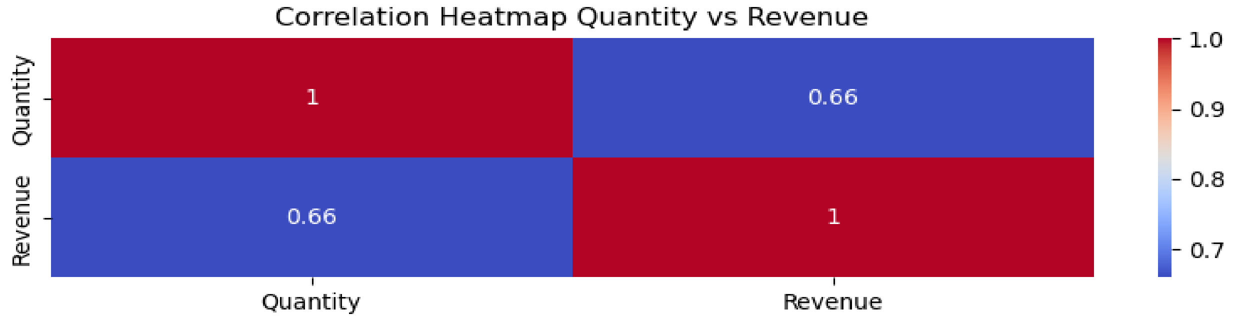


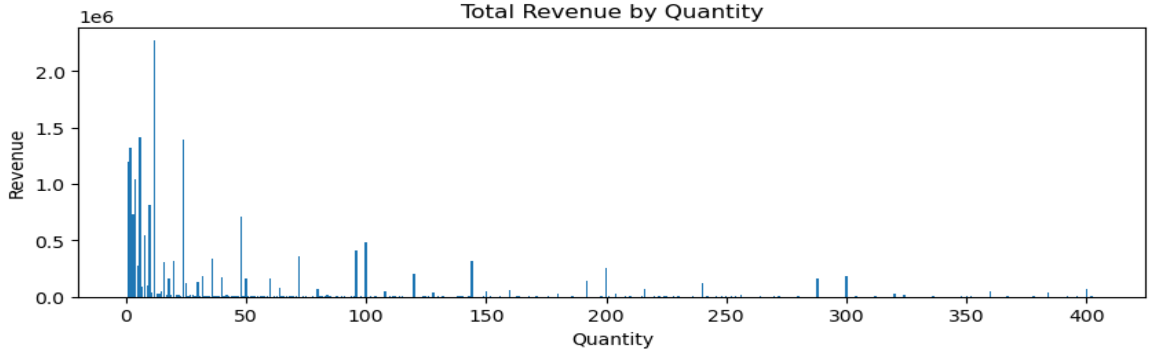




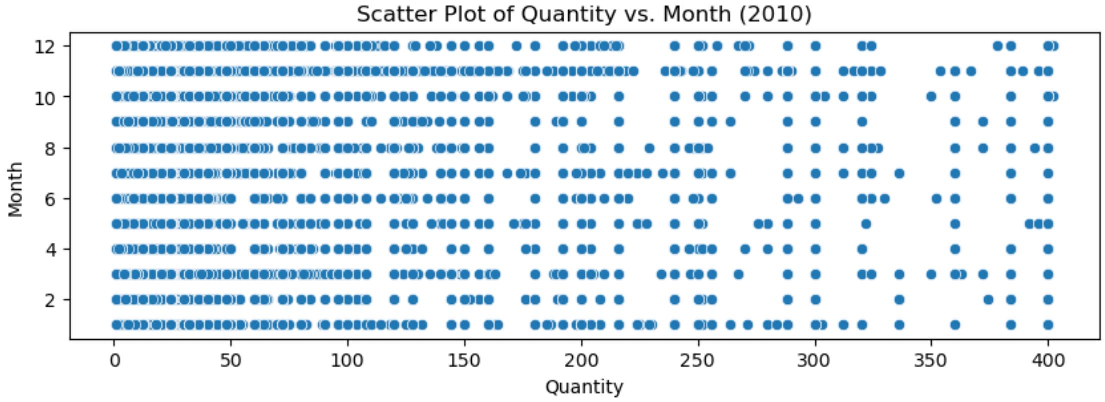
4.2.2 Quantity vs Revenue

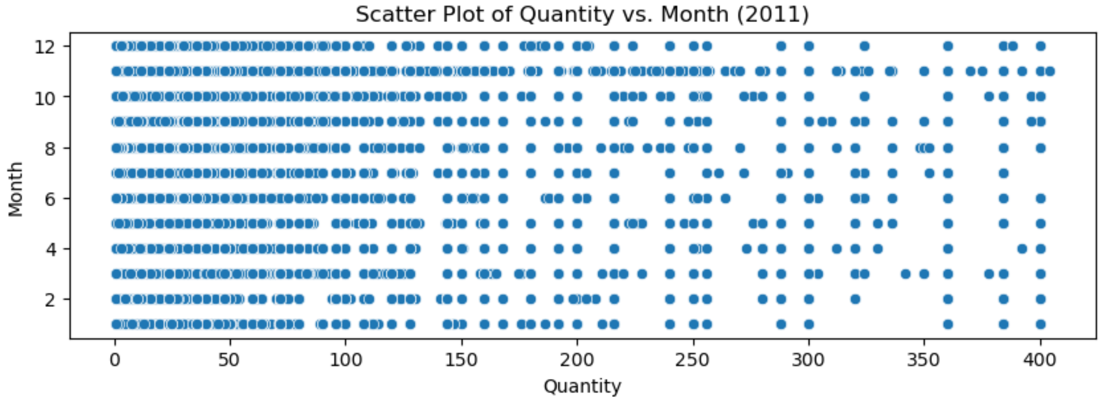


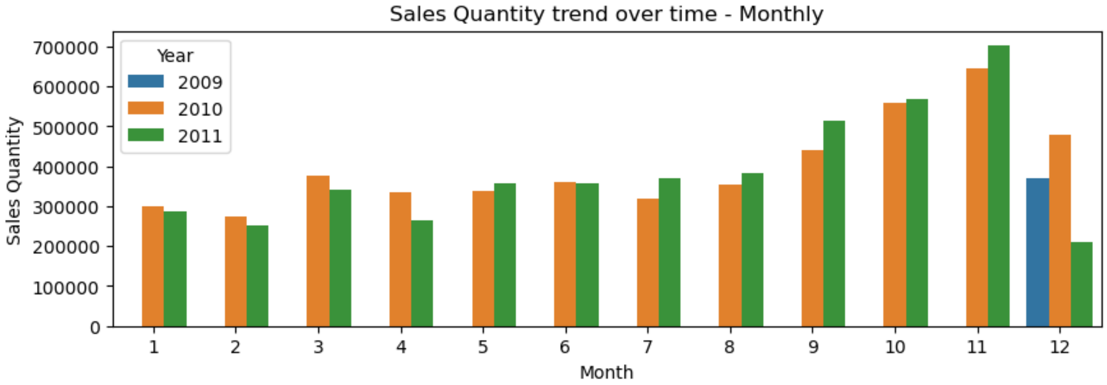


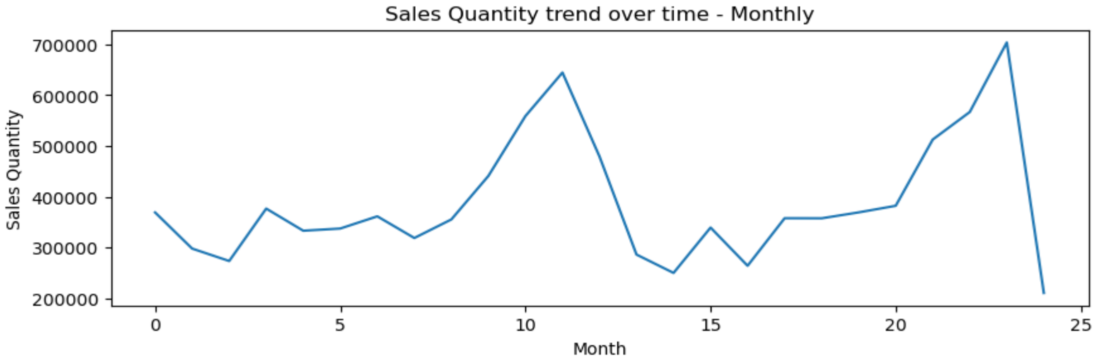


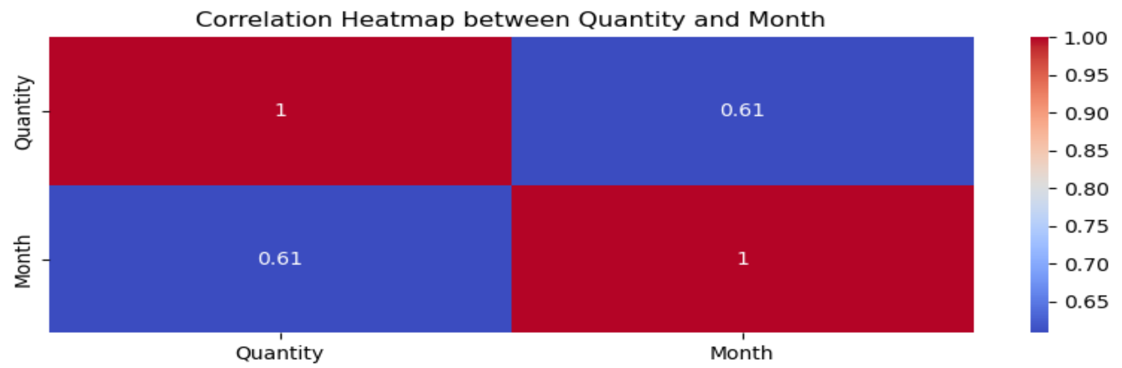
4.2.3 Quantity Vs Month


Scatter Plot of Quantity vs. Month (2010)

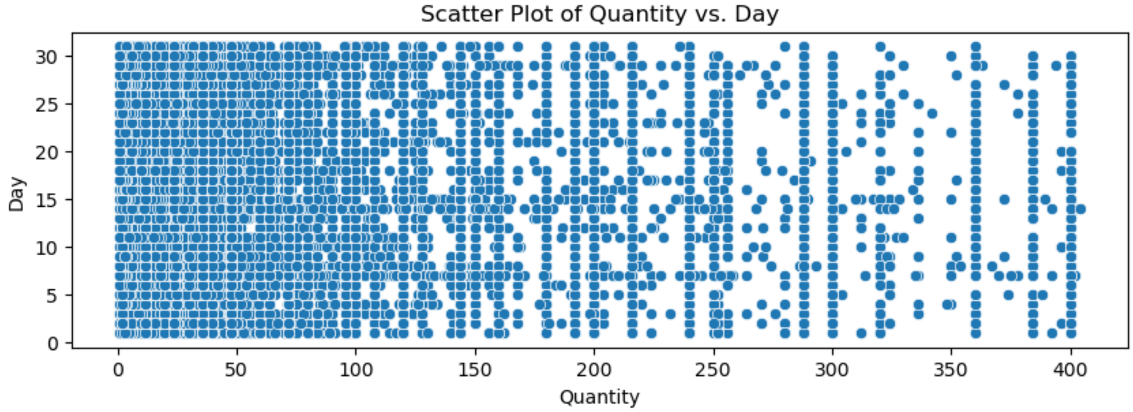

Scatter Plot of Quantity vs. Month (2011)

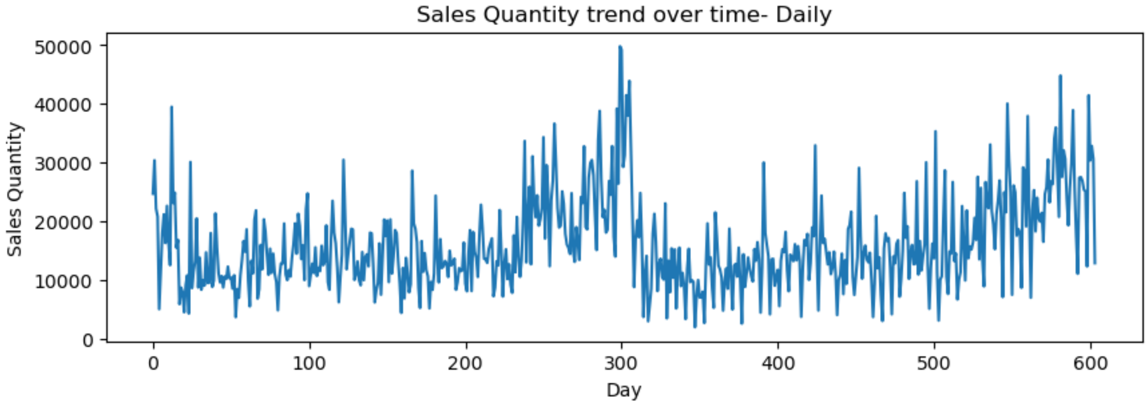

Sales Quantity trend over time - Monthly

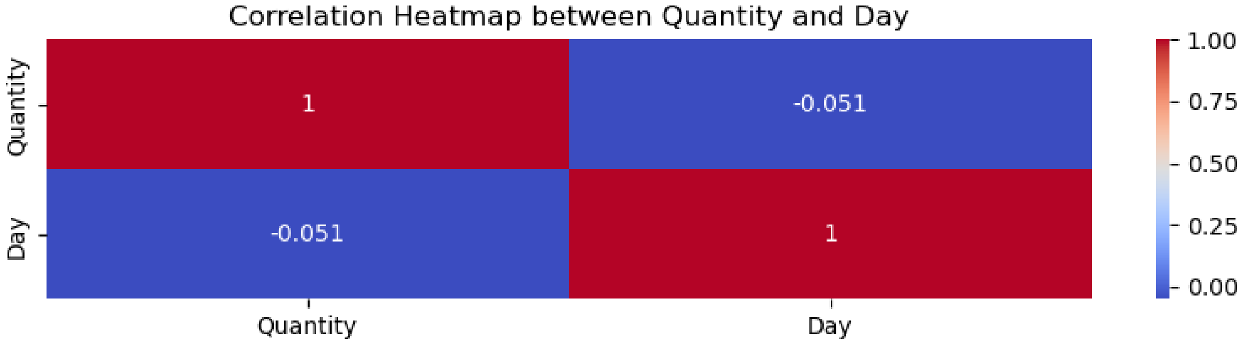

Sales Quantity trend over time - Monthly



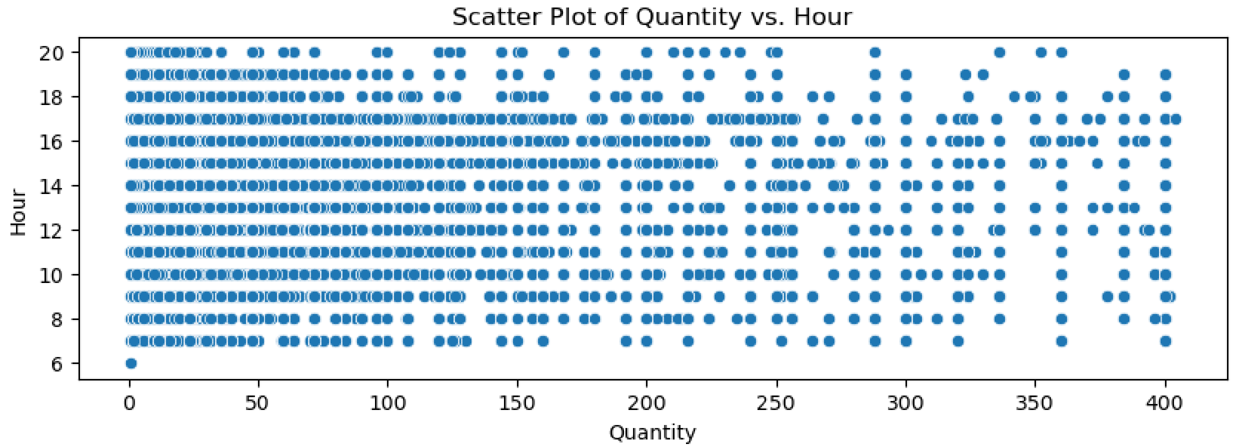
4.2.4 Quantity vs Day

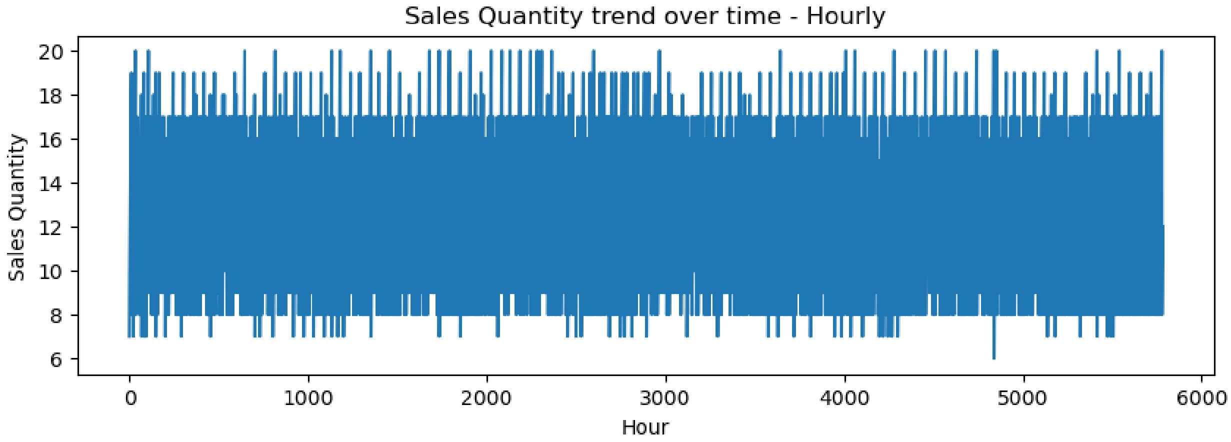


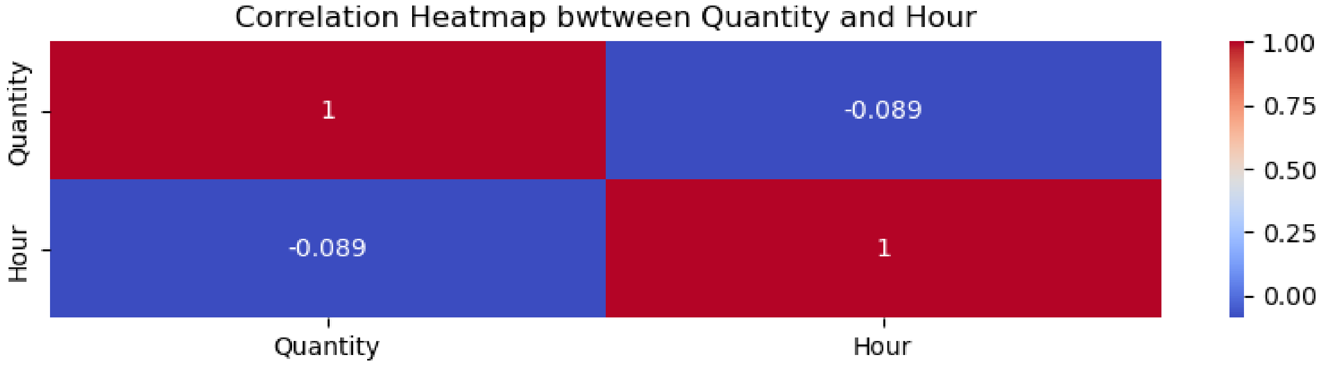




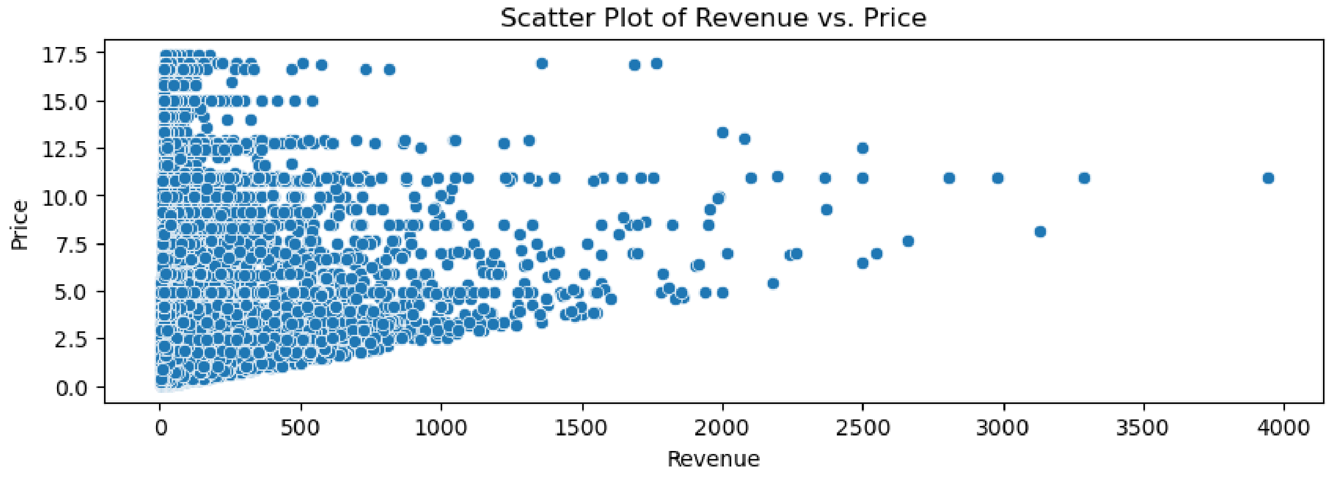
4.2.5 Quantity Vs Hour

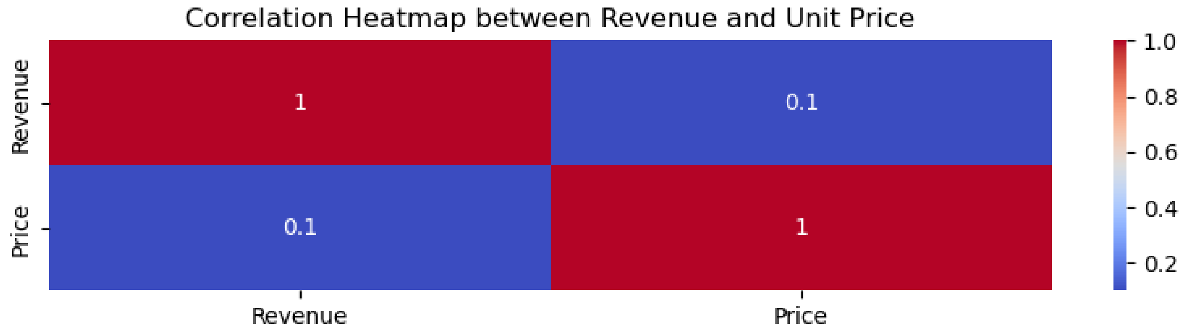


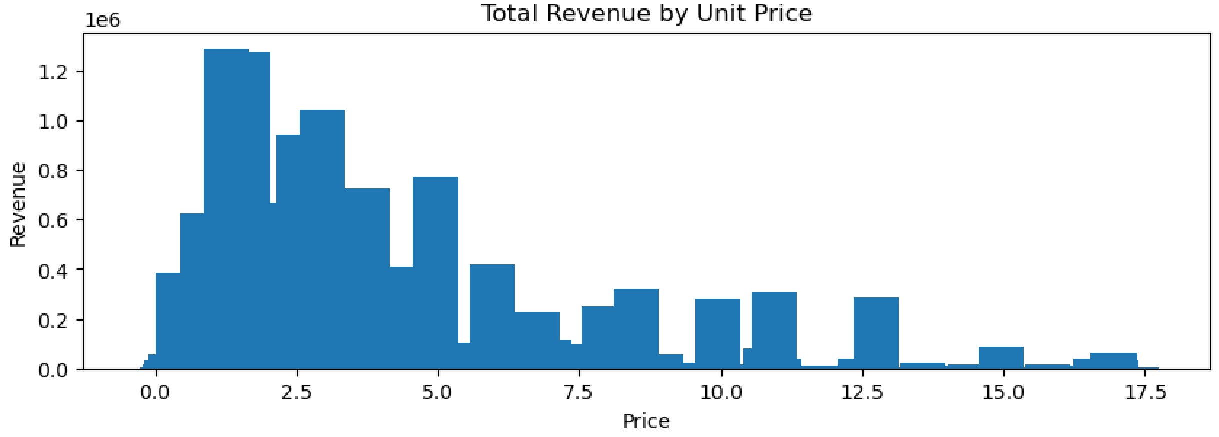




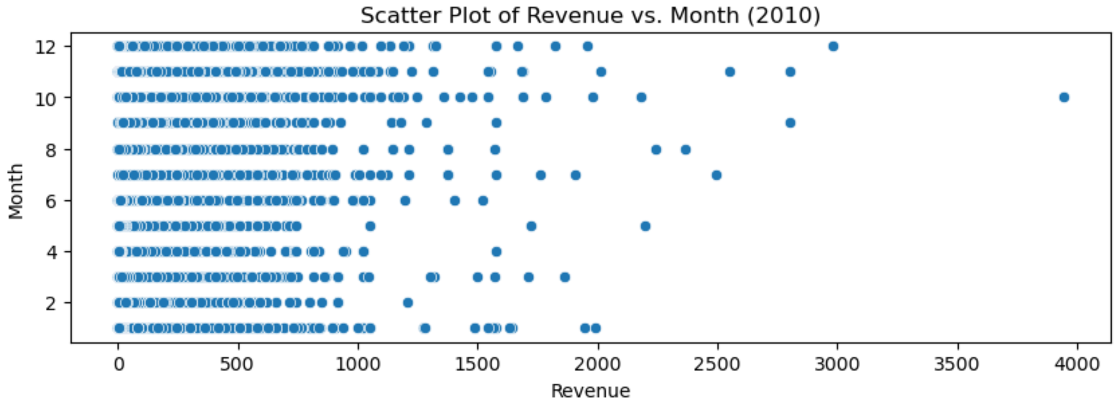
4.2.6 Revenue Vs Price

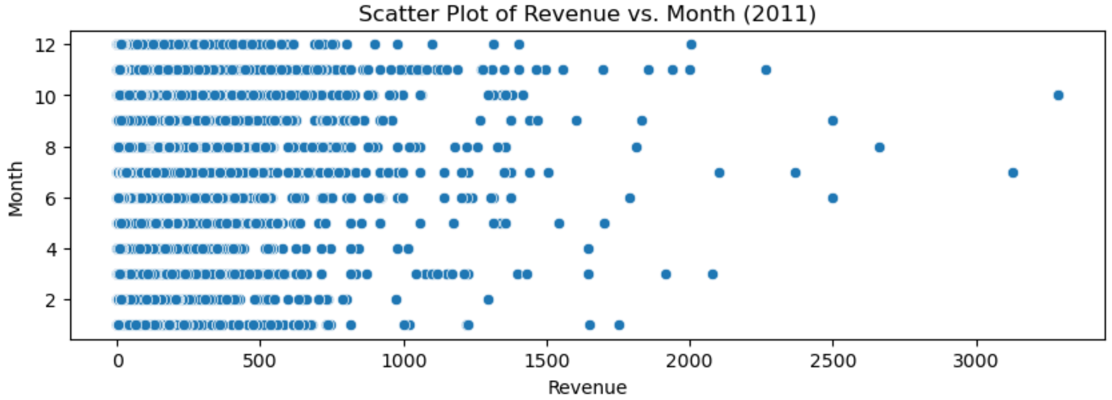


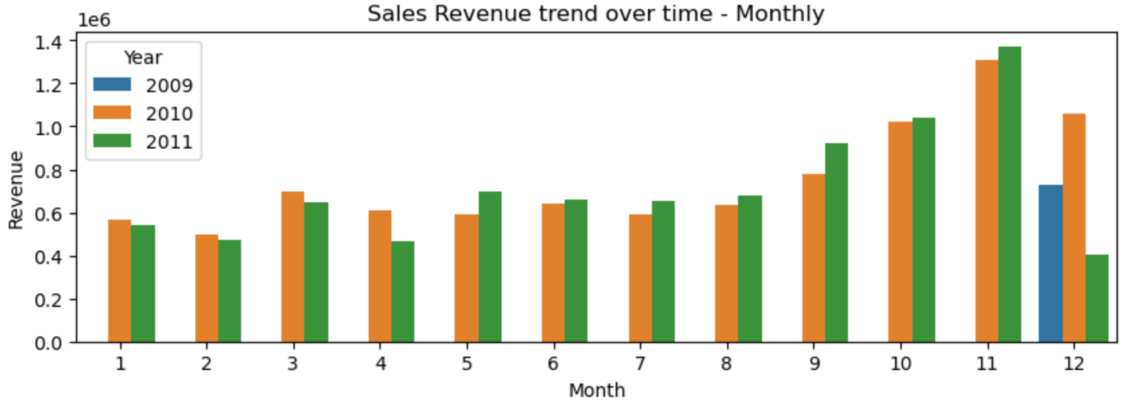


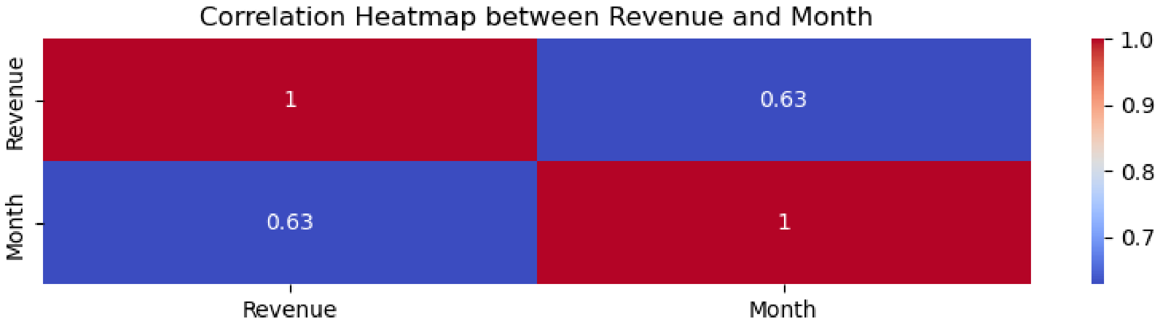


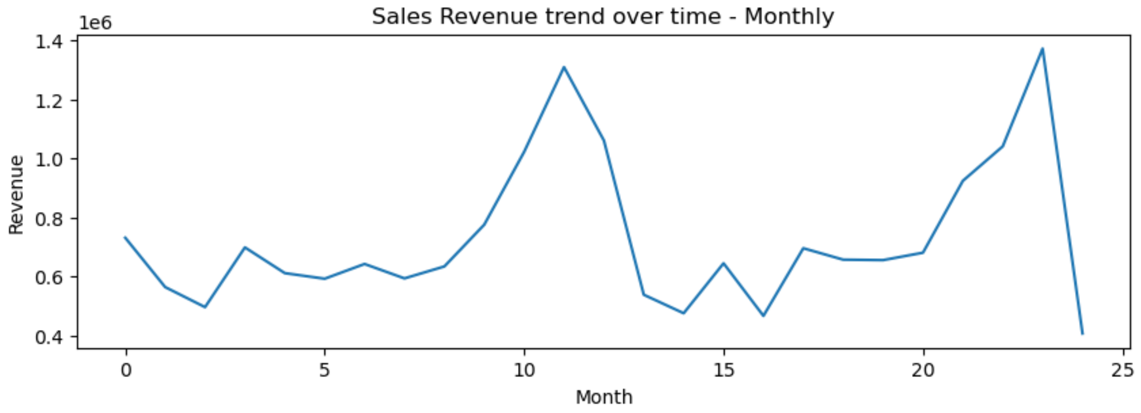
4.2.7 Revenue Vs Month


Scatter Plot of Revenue vs. Month (2010)

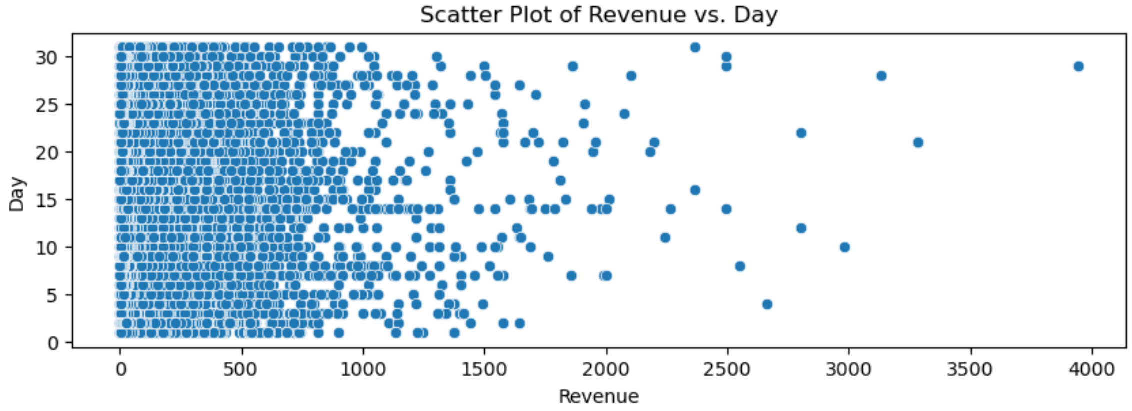

Scatter Plot of Revenue vs. Month (2011)

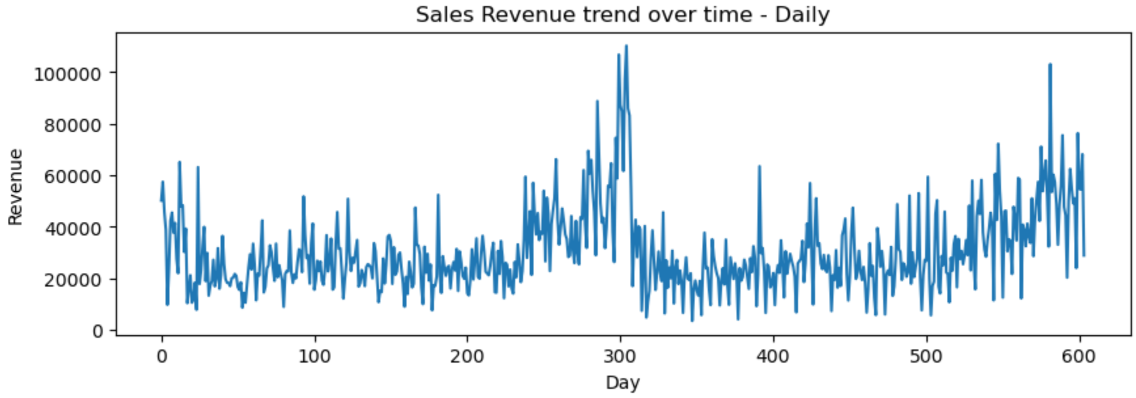

Sales Revenue trend over time - Monthly

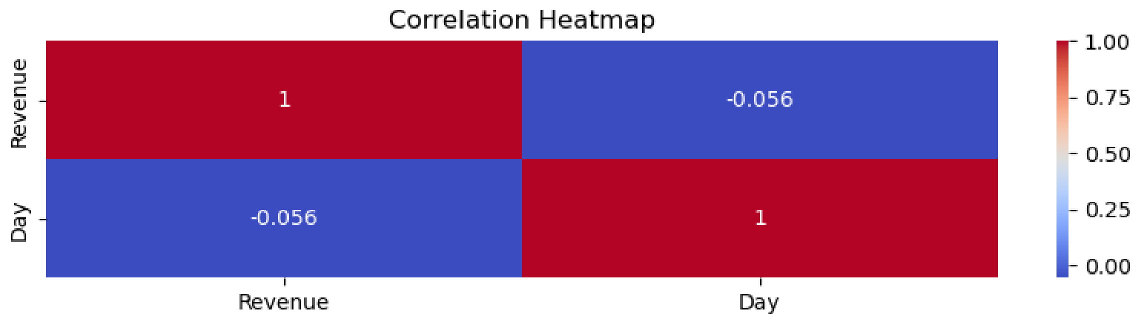

Correlation Heatmap between Revenue and Month



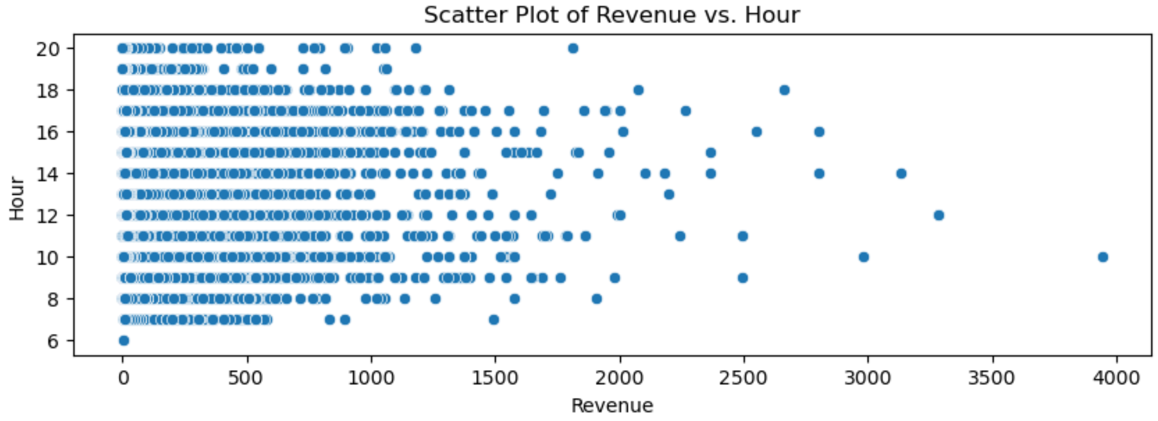
4.2.8 Revenue Vs Day

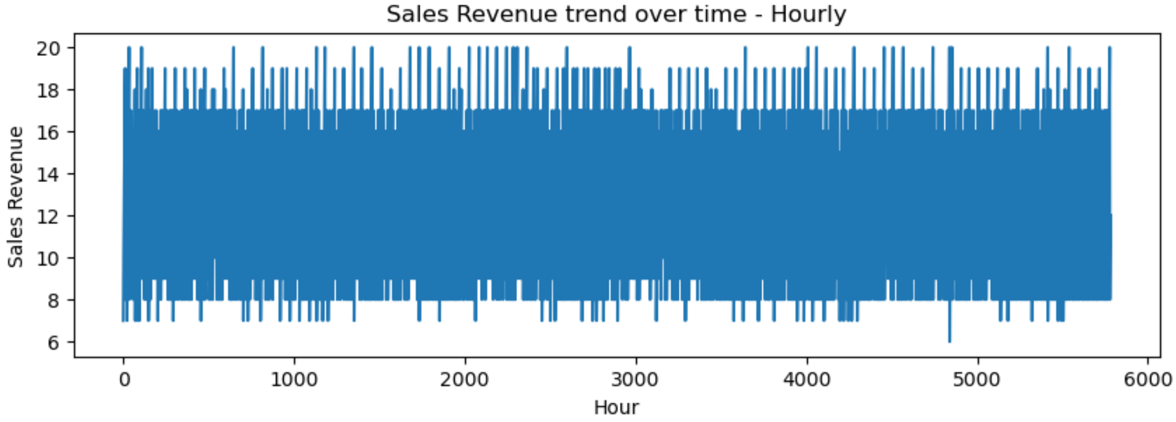


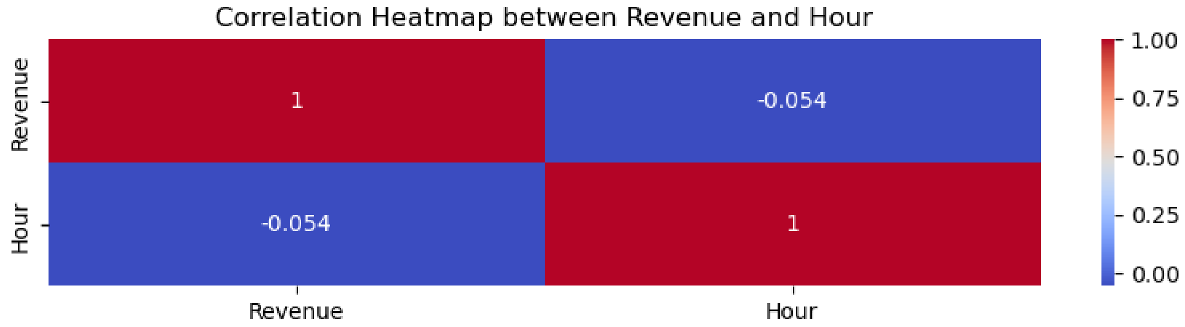




4.2.9 Revenue Vs Hour

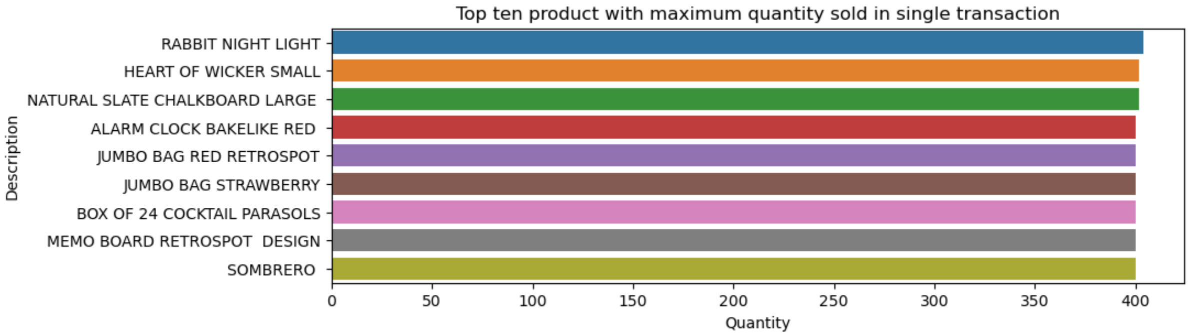


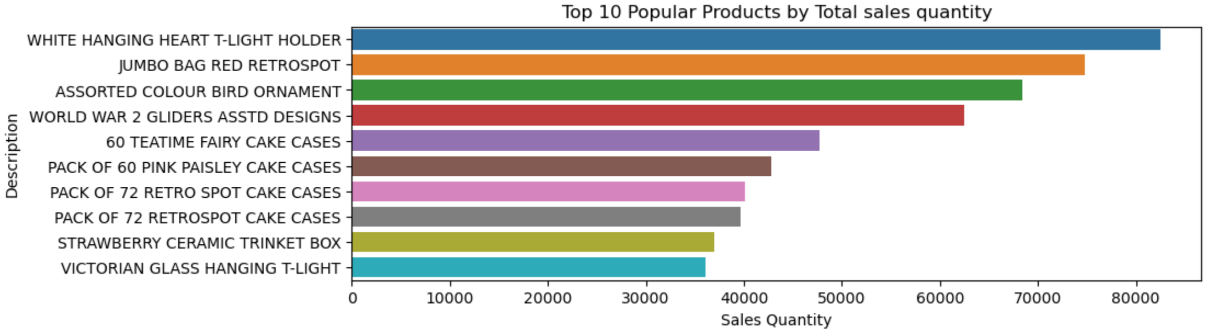




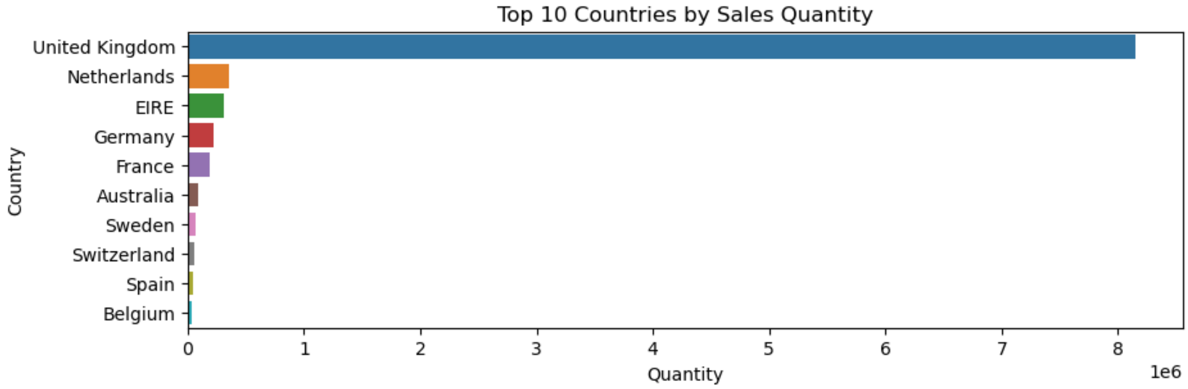
4.3 Bivariate Analysis for Numeric vs Nominal

4.3.1 Quantity vs Product Description

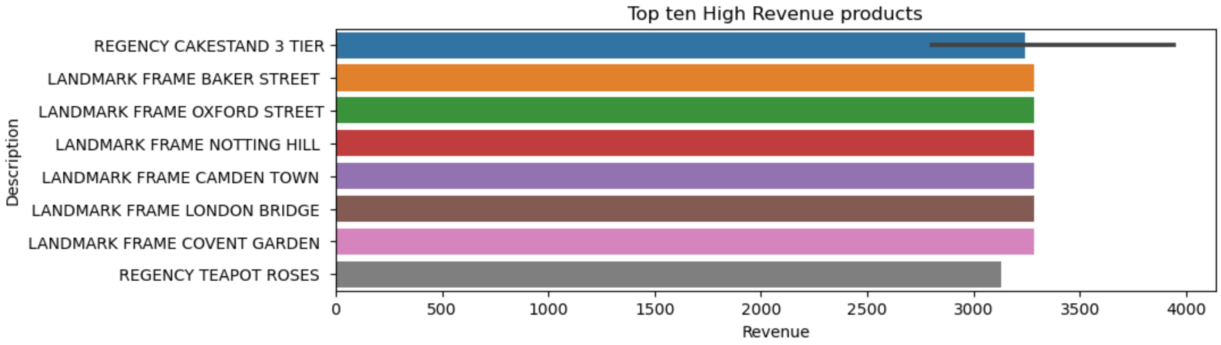


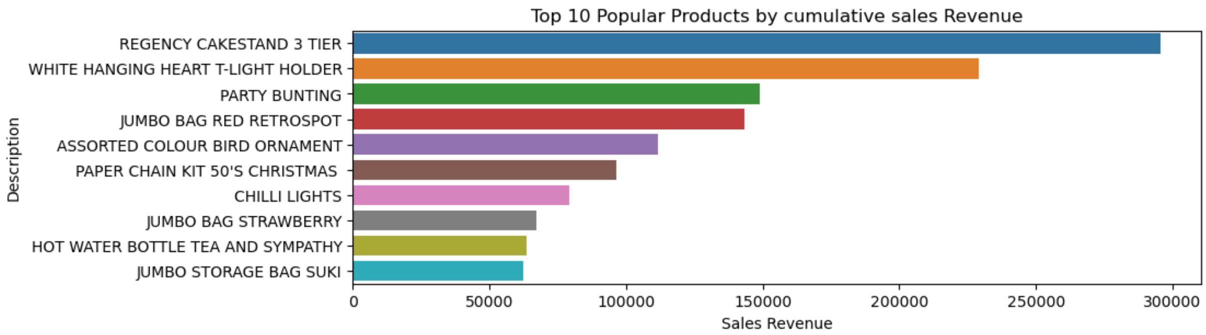


4.3.2 Quantity vs Country

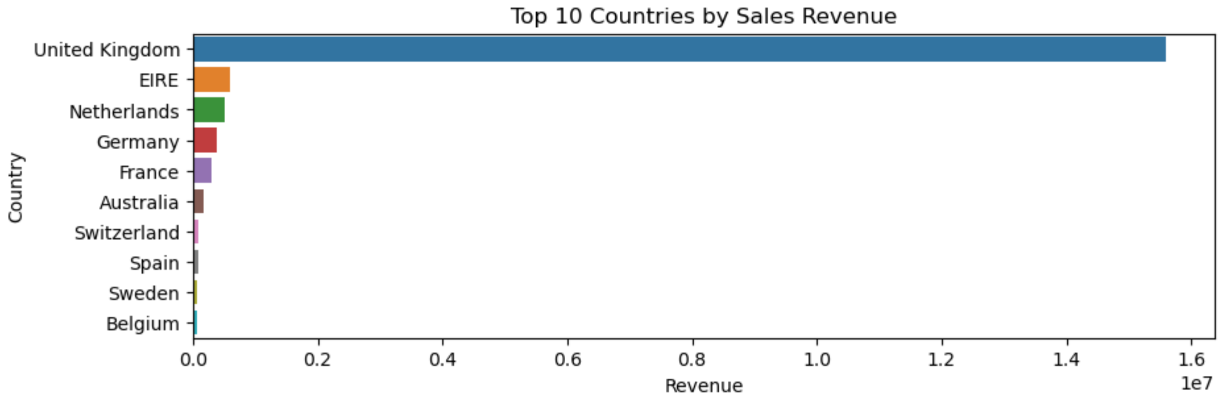


4.3.3 Revenue vs Product Description

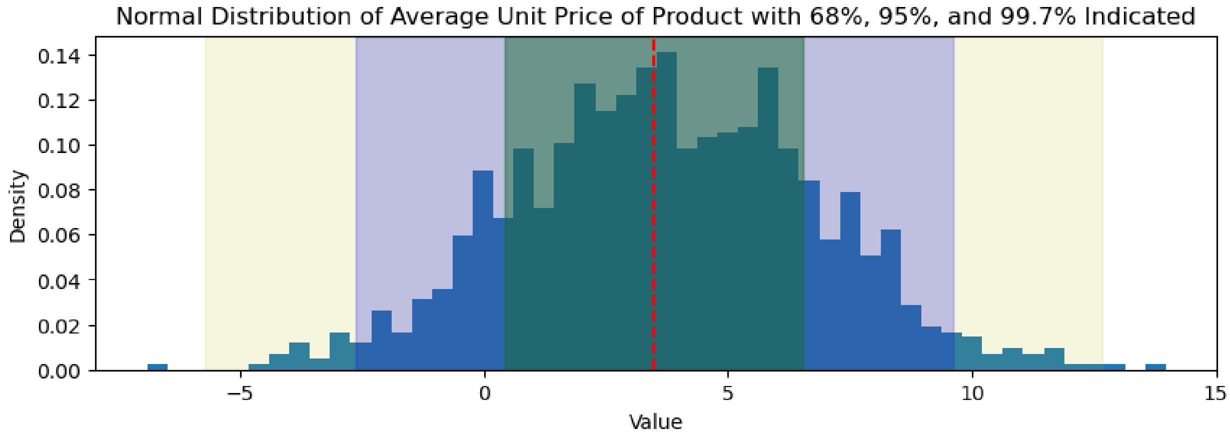


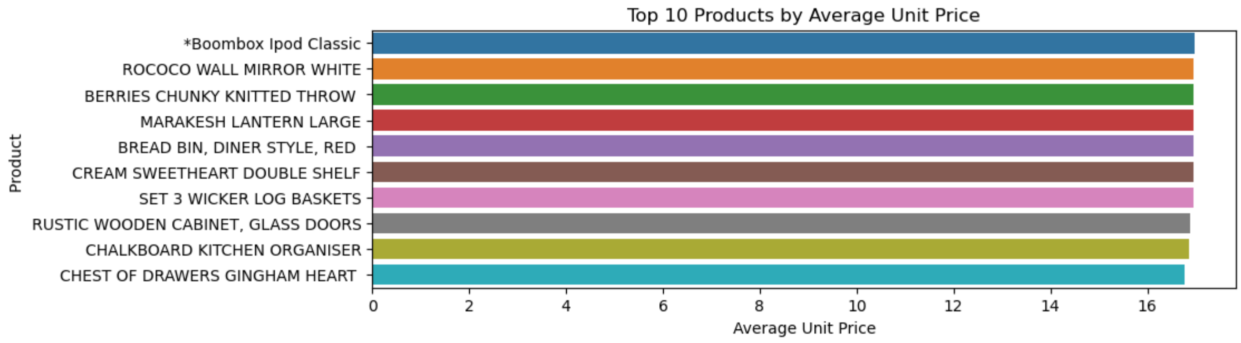


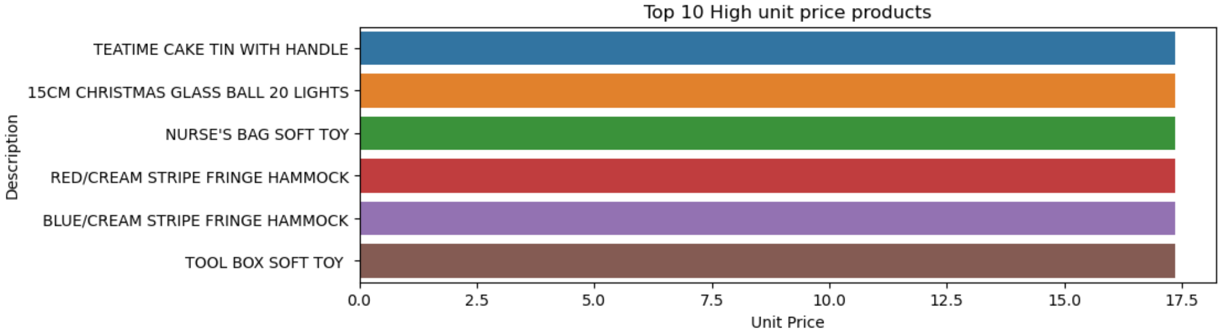
4.3.4 Revenue vs Country



4.3.5 Unit price vs Product Description

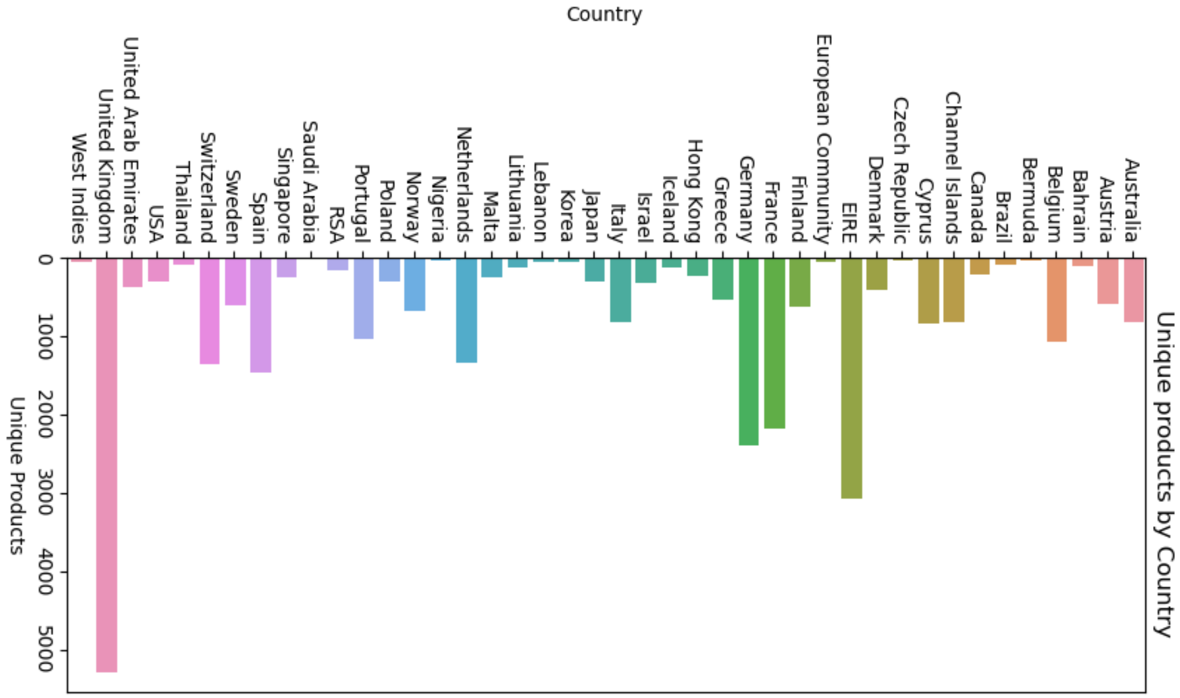


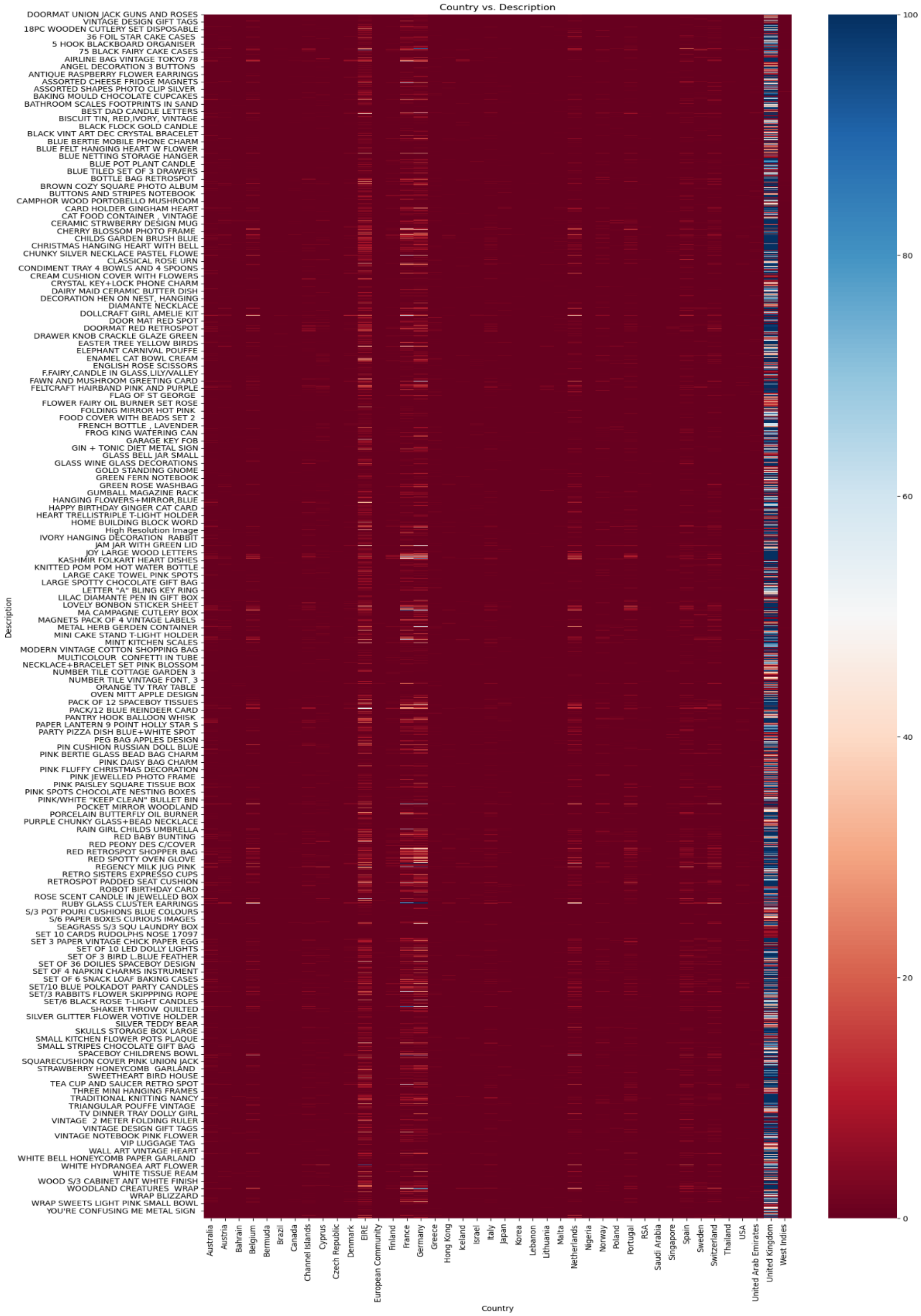




4.4 Bivariate Analysis for Nominal vs. Nominal

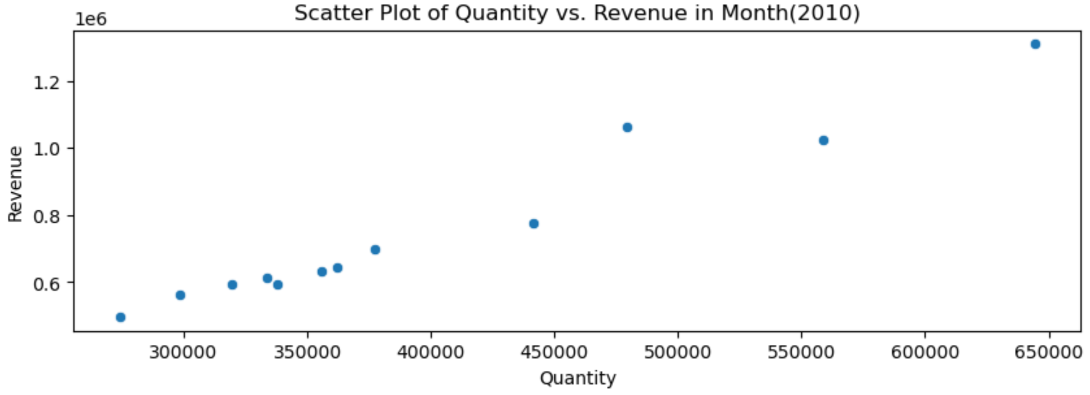
4.4.1 Country Vs Product Description

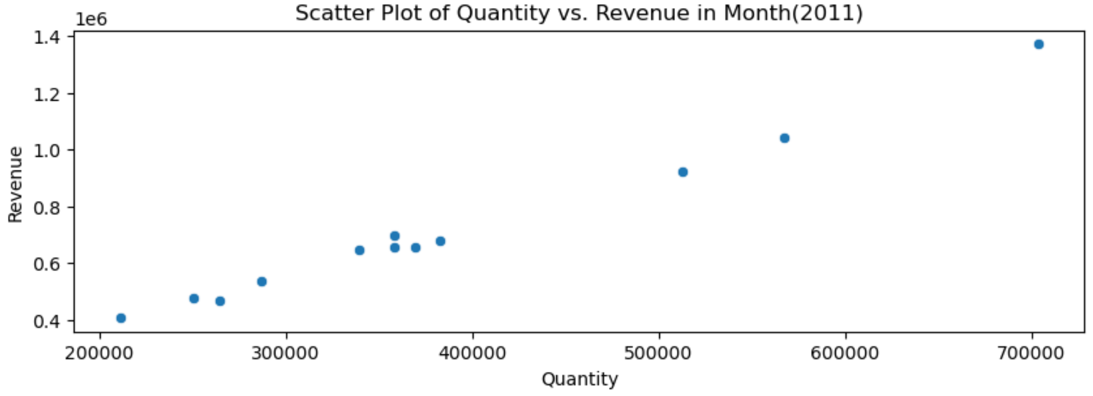


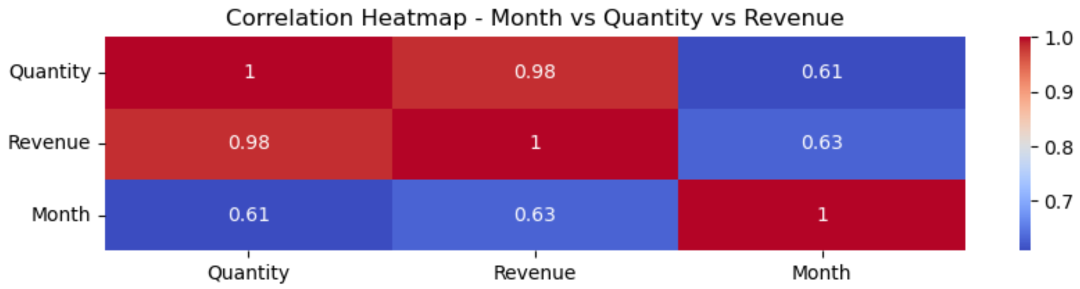

Country vs. Description
Description
DOORMAT UNION JACK GUNS AND ROSES
VINTAGE DESIGN GIFT TAGS
18PC WOODEN CUTLERY SET DISPOSABLE
36 FOIL STAR CAKE CASES
5 HOOK BLACKBOARD ORGANISER
75 BLACK FAIRY CAKE CASES
AIRLINE BAG VINTAGE TOKYO 78
ANGEL DECORATION 3 BUTTONS
ANTIQUE RASPBERRY FLOWER EARRINGS
ASSORTED CHEESE FRIDGE MAGNETS
ASSORTED SHAPES PHOTO CLIP SILVER
BAKING MOULD CHOCOLATE CUPCAKES
BATHROOM SCALES FOOTPRINTS IN SAND
BEST DAD CANDLE LETTERS
BISCUIT TIN, RED,IVORY, VINTAGE
BLACK FLOCK GOLD CANDLE
BLACK VINT ART DEC CRYSTAL BRACELET
BLUE BERTIE MOBILE PHONE CHARM
BLUE FELT HANGING HEART W FLOWER
BLUE NETTING STORAGE HANGER
BLUE POT PLANT CANDLE
BLUE TILED SET OF 3 DRAWERS
BOTTLE BAG RETROSPOT
BROWN COZY SQUARE PHOTO ALBUM
BUTTONS AND STRIPES NOTEBOOK
CAMPHOR WOOD PORTOBELLO MUSHROOM
CARD HOLDER GINGHAM HEART
CAT FOOD CONTAINER , VINTAGE
CERAMIC STRWBERRY DESIGN MUG
CHERRY BLOSSOM PHOTO FRAME
CHILDS GARDEN BRUSH BLUE
CHRISTMAS HANGING HEART WITH BELL
CHUNKY SILVER NECKLACE PASTEL FLOWE
CLASSICAL ROSE URN
CONDIMENT TRAY 4 BOWLS AND 4 SPOONS
CREAM CUSHION COVER WITH FLOWERS
CRYSTAL KEY+LOCK PHONE CHARM
DAIRY MAID CERAMIC BUTTER DISH
DECORATION HEN ON NEST, HANGING
DIAMANTE NECKLACE
DOLLCRAFT GIRL AMELIE KIT
DOOR MAT RED SPOT
DOORMAT RED RETROSPOT
DRAWER KNOB CRACKLE GLAZE GREEN
EASTER TREE YELLOW BIRDS
ELEPHANT CARNIVAL POUFFE
ENAMEL CAT BOWL CREAM
ENGLISH ROSE SCISSORS
F.FAIRY,CANDLE IN GLASS,LILY/VALLEY
FAWN AND MUSHROOM GREETING CARD
FELTCRAFT HAIRBAND PINK AND PURPLE
FLAG OF ST GEORGE
FLOWER FAIRY OIL BURNER SET ROSE
FOLDING MIRROR HOT PINK
FOOD COVER WITH BEADS SET 2
FRENCH BOTTLE , LAVENDER
FROG KING WATERING CAN
GARAGE KEY FOB
GIN + TONIC DIET METAL SIGN
GLASS BELL JAR SMALL
GLASS WINE GLASS DECORATIONS
GOLD STANDING GNOME
GREEN FERN NOTEBOOK
GREEN ROSE WASHBAG
GUMBALL MAGAZINE RACK
HANGING FLOWERS+MIRROR,BLUE
HAPPY BIRTHDAY GINGER CAT CARD
HEART TRELLISTRIPLE T-LIGHT HOLDER
HOME BUILDING BLOCK WORD
High Resolution Image
IVORY HANGING DECORATION RABBIT
JAM JAR WITH GREEN LID
JOY LARGE WOOD LETTERS
KASHMIR FOLKART HEART DISHES
KNITTED POM POM HOT WATER BOTTLE
LARGE CAKE TOWEL PINK SPOTS
LARGE SPOTTY CHOCOLATE GIFT BAG
LETTER "A" BLING KEY RING
LILAC DIAMANTE PEN IN GIFT BOX
LOVELY BONBON STICKER SHEET
MA CAMPAGNE CUTLERY BOX
MAGNETS PACK OF 4 VINTAGE LABELS
METAL HERB GERDEN CONTAINER
MINI CAKE STAND T-LIGHT HOLDER
MINT KITCHEN SCALES
MODERN VINTAGE COTTON SHOPPING BAG
MULTICOLOUR CONFETTI IN TUBE
NECKLACE+BRACELET SET PINK BLOSSOM
NUMBER TILE COTTAGE GARDEN 3
NUMBER TILE VINTAGE FONT, 3
ORANGE TV TRAY TABLE
OVEN MITT APPLE DESIGN
PACK OF 12 SPACEBOY TISSUES
PACK/12 BLUE REINDEER CARD
PANTRY HOOK BALLOON WHISK
PAPER LANTERN 9 POINT HOLLY STAR S
PARTY PIZZA DISH BLUE+WHITE SPOT
PEG BAG APPLES DESIGN
PIN CUSHION RUSSIAN DOLL BLUE
PINK BERTIE GLASS BEAD BAG CHARM
PINK DAISY BAG CHARM
PINK FLUFFY CHRISTMAS DECORATION
PINK JEWELLED PHOTO FRAME
PINK PAISLEY SQUARE TISSUE BOX
PINK SPOTS CHOCOLATE NESTING BOXES
PINK/WHITE "KEEP CLEAN" BULLET BIN
POCKET MIRROR WOODLAND
PORCELAIN BUTTERFLY OIL BURNER
PURPLE CHUNKY GLASS+BEAD NECKLACE
RAIN GIRL CHILDS UMBRELLA
RED BABY BUNTING
RED PEONY DES C/COVER
RED RETROSPOT SHOPPER BAG
RED SPOTTY OVEN GLOVE
REGENCY MILK JUG PINK
RETRO SISTERS EXPRESSO CUPS
RETROSPOT PADDED SEAT CUSHION
ROBOT BIRTHDAY CARD
ROSE SCENT CANDLE IN JEWELLED BOX
RUBY GLASS CLUSTER EARRINGS
S/3 POT POURI CUSHIONS BLUE COLOURS
S/6 PAPER BOXES CURIOUS IMAGES
SEAGRASS S/3 SQU LAUNDRY BOX
SET 10 CARDS RUDOLPHS NOSE 17097
SET 3 PAPER VINTAGE CHICK PAPER EGG
SET OF 10 LED DOLLY LIGHTS
SET OF 3 BIRD L.BLUE FEATHER
SET OF 36 DOILIES SPACEBOY DESIGN
SET OF 4 NAPKIN CHARMS INSTRUMENT
SET OF 6 SNACK LOAF BAKING CASES
SET/10 BLUE POLKADOT PARTY CANDLES
SET/3 RABBITS FLOWER SKIPPPING ROPE
SET/6 BLACK ROSE T-LIGHT CANDLES
SHAKER THROW QUILTED
SILVER GLITTER FLOWER VOTIVE HOLDER
SILVER TEDDY BEAR
SKULLS STORAGE BOX LARGE
SMALL KITCHEN FLOWER POTS PLAQUE
SMALL STRIPES CHOCOLATE GIFT BAG
SPACEBOY CHILDRENS BOWL
SQUARECUSHION COVER PINK UNION JACK
STRAWBERRY HONEYCOMB GARLAND
SWEETHEART BIRD HOUSE
TEA CUP AND SAUCER RETRO SPOT
THREE MINI HANGING FRAMES
TRADITIONAL KNITTING NANCY
TRIANGULAR POUFFE VINTAGE
TV DINNER TRAY DOLLY GIRL
VINTAGE 2 METER FOLDING RULER
VINTAGE DESIGN GIFT TAGS
VINTAGE NOTEBOOK PINK FLOWER
VIP LUGGAGE TAG
WALL ART VINTAGE HEART
WHITE BELL HONEYCOMB PAPER GARLAND
WHITE HYDRANGEA ART FLOWER
WHITE TISSUE REAM
WOOD S/3 CABINET ANT WHITE FINISH
WOODLAND CREATURES WRAP
WRAP BLIZZARD
WRAP SWEETS LIGHT PINK SMALL BOWL
YOU'RE CONFUSING ME METAL SIGN
Country
Australia
Austria
Bahrain
Belgium
Bermuda
Brazil
Canada
Channel Islands
Cyprus
Czech Republic
Denmark
EIRE
European Community
Finland
France
Germany
Greece
Hong Kong
Iceland
Israel
Italy
Japan
Korea
Lebanon
Lithuania
Malta
Netherlands
Nigeria
Norway
Poland
Portugal
RSA
Saudi Arabia
Singapore
Spain
Sweden
Switzerland
Thailand
USA
United Arab Emirates
United Kingdom
West Indies
100
80
60
40
20
0

4.5 Multivariate Analysis

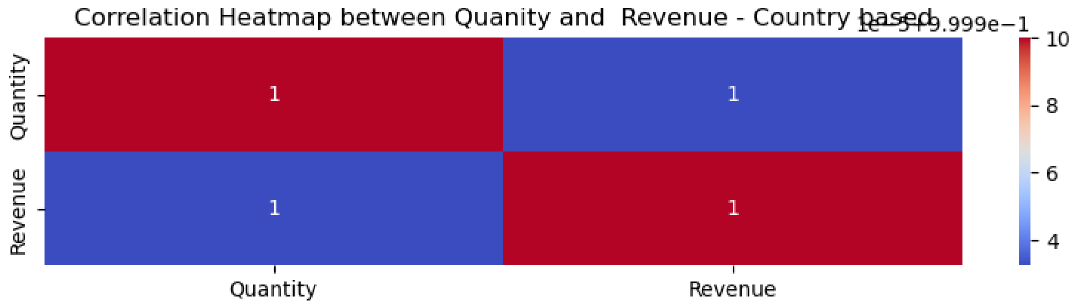
4.5.1 Month – Quantity – Revenue

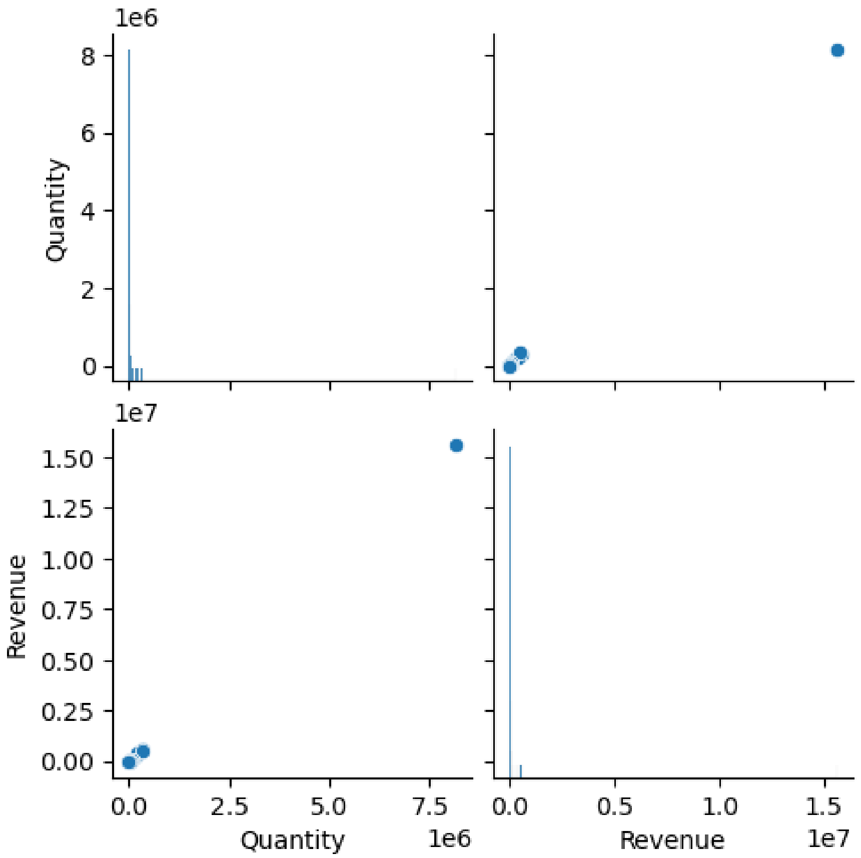




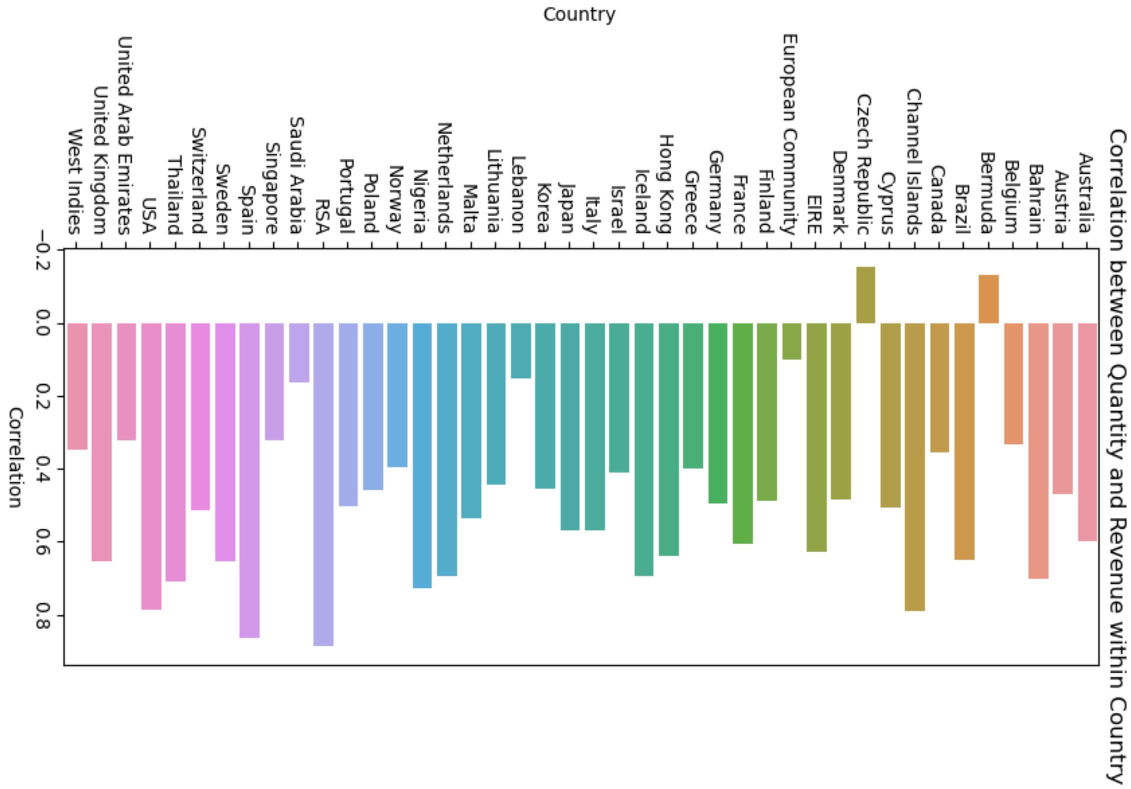


4.5.2 Country – Quantity – Revenue



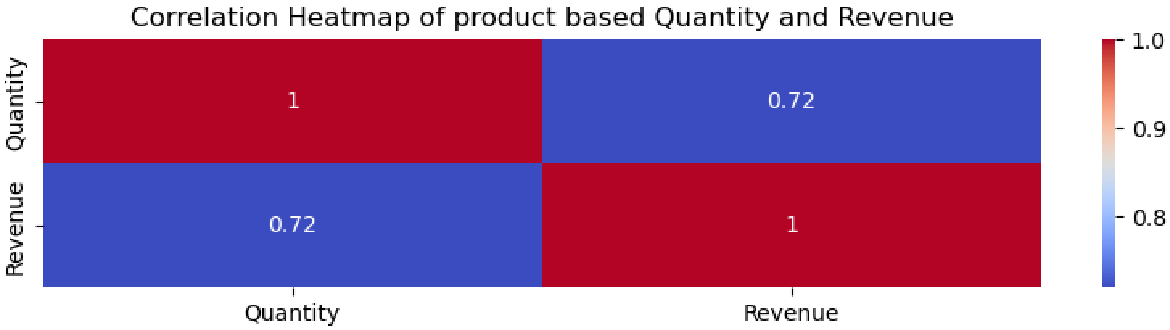


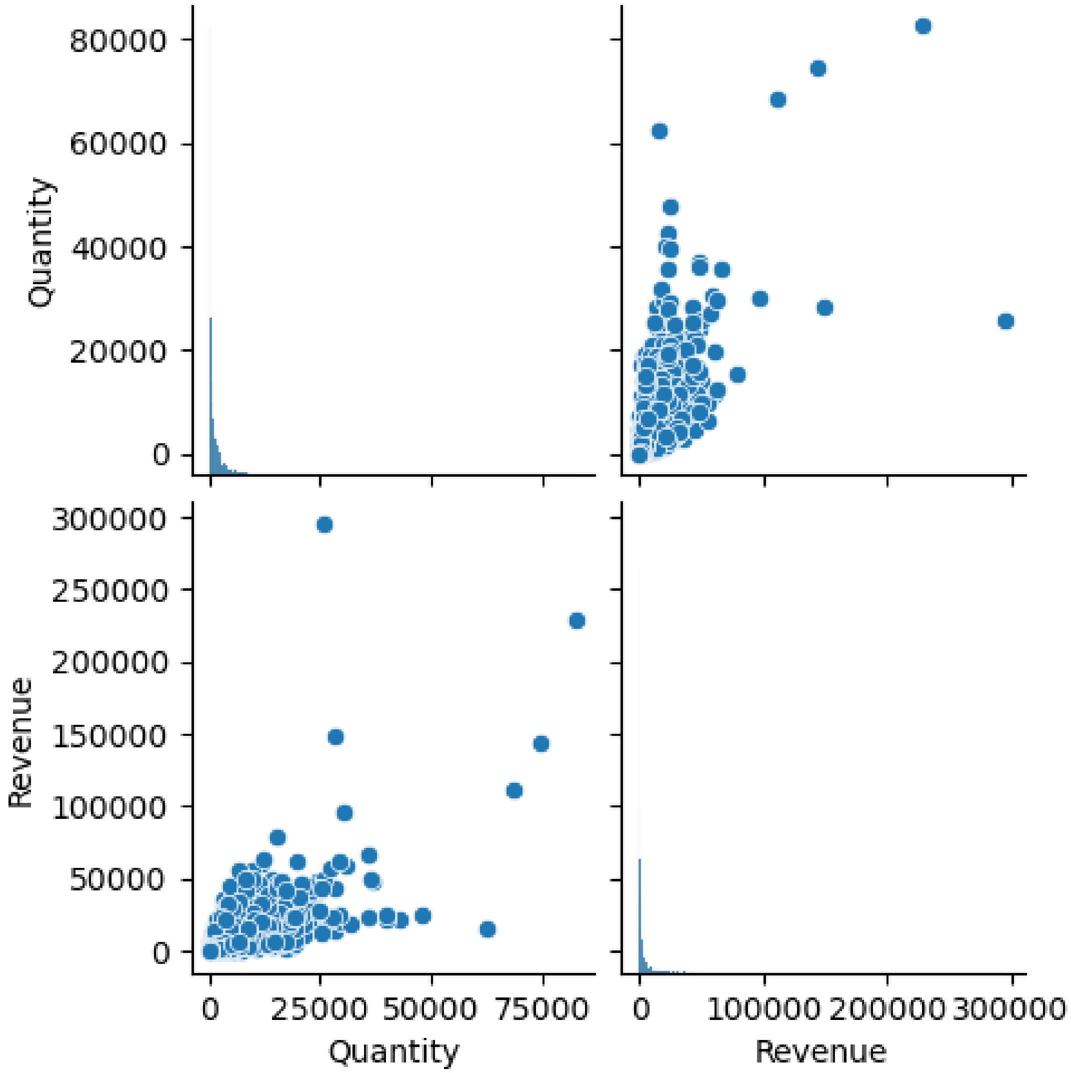
1e6
8
6
4
2
0
Quantity
1e7
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Revenue
0.0
2.5
5.0
7.5
Quantity
1e6
0.0
0.5
1.0
1.5
Revenue
1e7



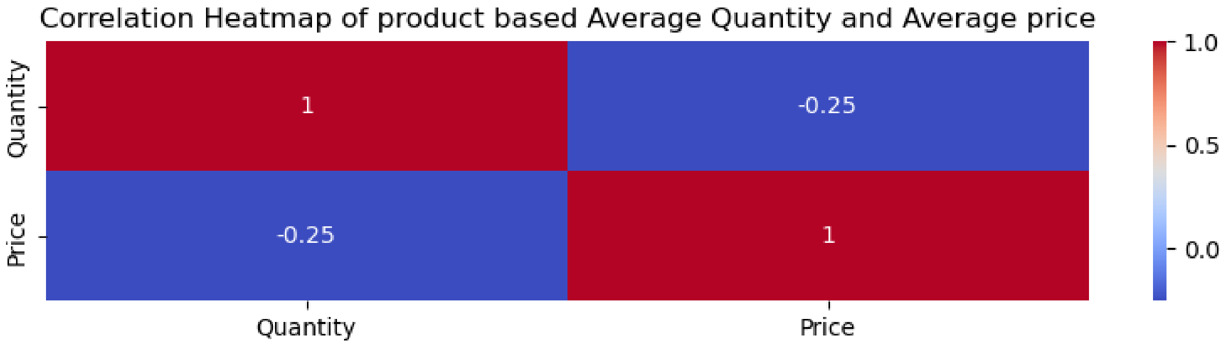
Country
West Indies
United Kingdom
United Arab Emirates
USA
Thailand
Switzerland
Sweden
Spain
Singapore
Saudi Arabia
RSA
Portugal
Poland
Norway
Nigeria
Netherlands
Malta
Lithuania
Lebanon
Korea
Japan
Italy
Israel
Iceland
Hong Kong
Greece
Germany
France
Finland
European Community
EIRE
Denmark
Czech Republic
Cyprus
Channel Islands
Canada
Brazil
Bermuda
Belgium
Bahrain
Austria
Australia
Correlation
−0.2
0.0
0.2
0.4
0.6
0.8
Correlation between Quantity and Revenue within Country

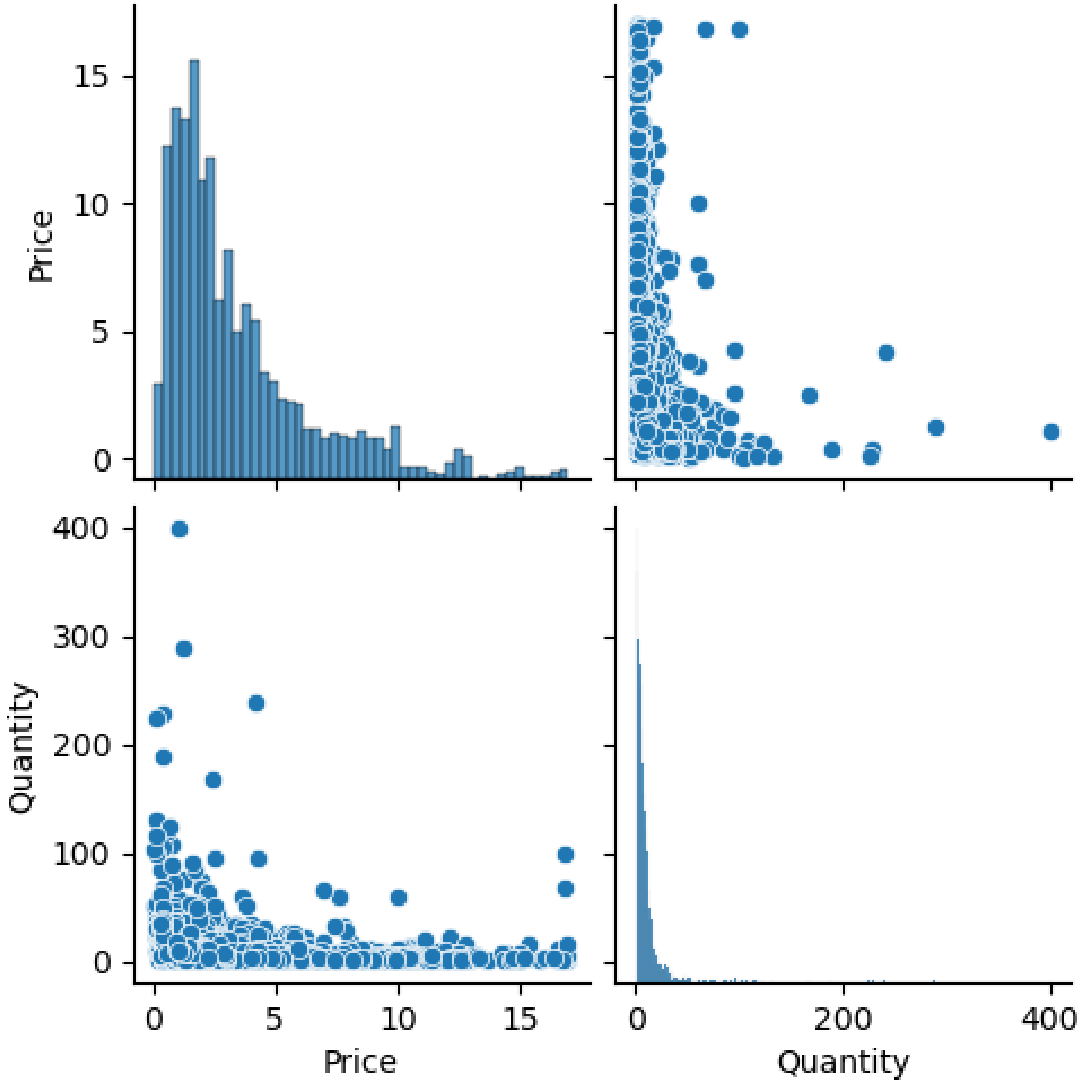
4.5.3 Product – Quantity – Revenue



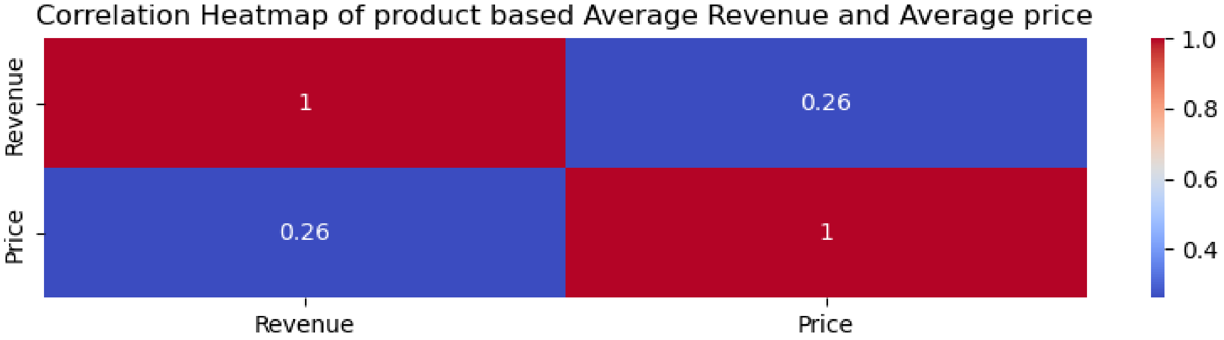


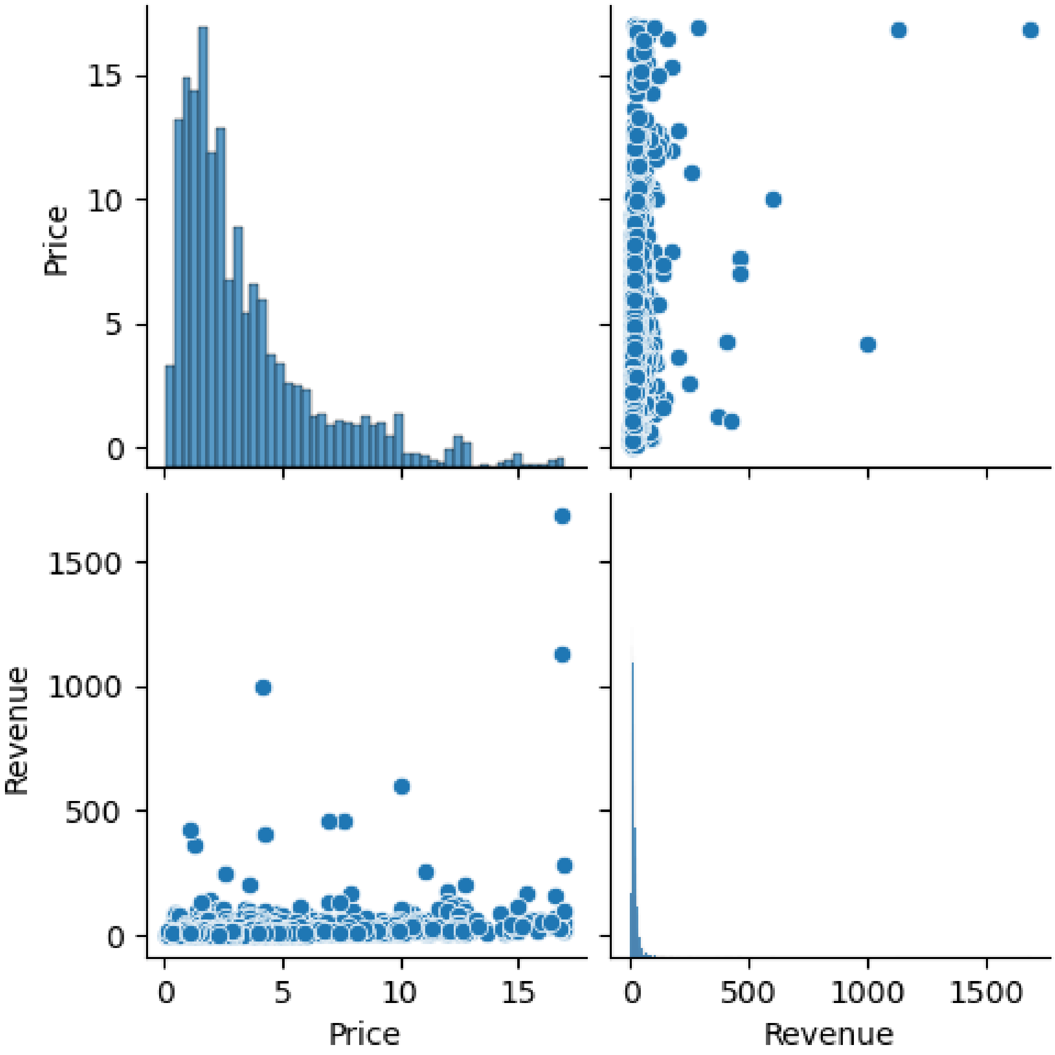
4.5.4 Product – Average Quantity – Average Unit Price



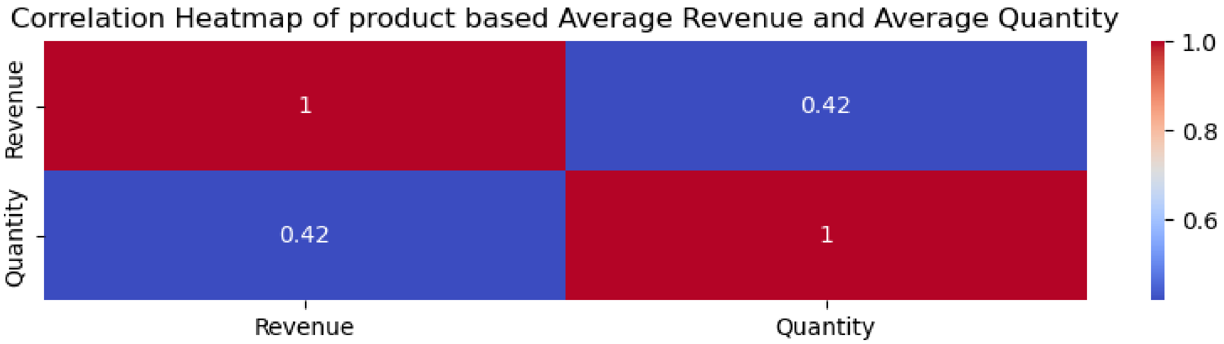


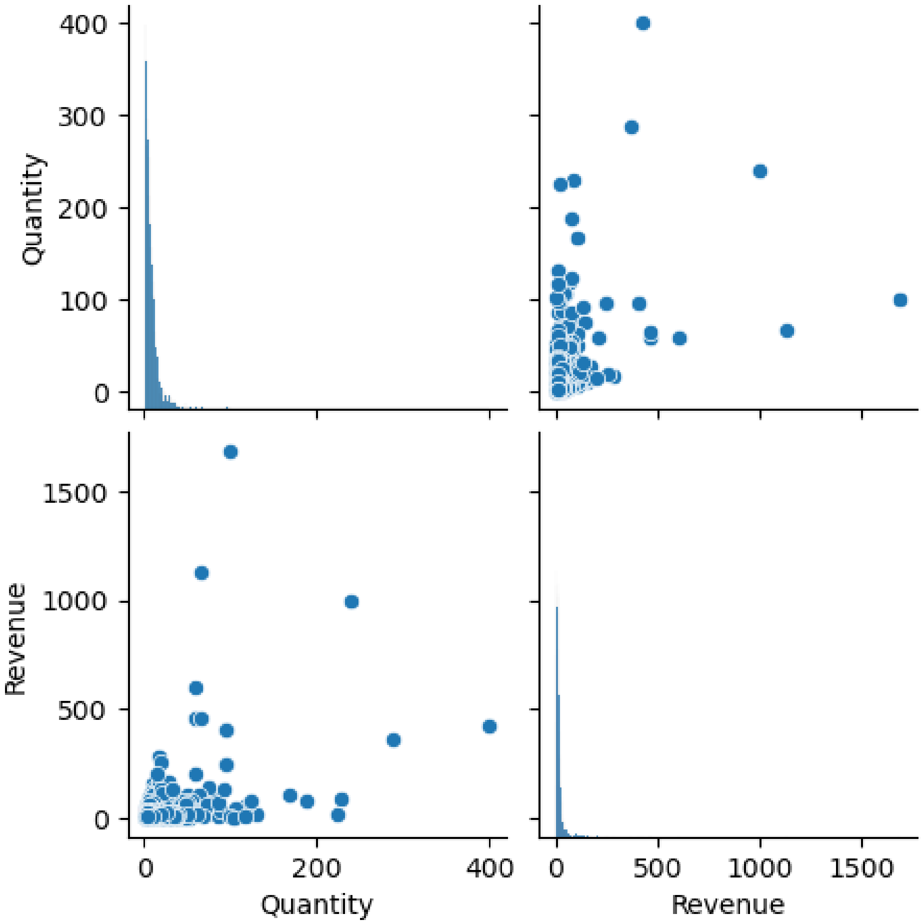
4.5.5 Product - Av Unit Price - Av Revenue





4.5.6 Product – Av Quantity – Av Revenue





5 Key Findings and Insights

1. Approximately 75% of goods were sold in quantities under 11 units. With a 0.61 correlation coefficient, product sales and month are strongly correlated. With a correlation of 0.66, product sales and revenue are linked. Quantity sold does not correlate with date (0.-051). No correlation exists between hours worked and product sales (0.-089).

2. 75% of products cost less than £4.13 per unit. Product unit prices depend on sales volume. Quantity and unit price have a weak negative correlation of -0.19. The number sold may somewhat decrease when the unit price rises. A 0.10 modest positive correlation exists between unit pricing and revenue.

3. About 75% of products had revenue under £17. Revenue and month have a positive correlation value of 0.63. Revenue and monthday are unrelated (0.-056). There is no correlation (-0.054) between revenue and hours worked.

4. The United Kingdom leads the market with 83% of total sales and 85% of total income. After the UK, Ireland, the Netherlands, Germany, and France have strong sales and revenue. Most countries, except Bermuda and the Czech Republic, have a strong positive link between quantity and revenue.

5. Monthly, a high positive correlation exists between quantity and revenue, with a value of 0.98. Revenue and quantity rose from January to November but fell in December. November, October, and September had the most sales in 2010 and 2011. Sales were steady throughout 2010 and 2011, but sales rose in 2011. Daily transactions occur throughout the month from 7 a.m. to 8 p.m. Peak activity is usually between 10 a.m. and 5 p.m.

6. There is a high positive correlation between total quantity and total revenue on product basis, with a value of 0.72. This means that revenue tends to rise over products, quantity. Top products by sales volume, revenue, and quantity: WHITE HANGING HEART T-LIGHT HOLDER [Unit price – 1.09 – 6.77], JUMBO BAG RED RETROSPOT [Unit

price – 1.65 – 5. 06], ASSORTED COLOUR BIRD ORNAMENT [Unit price – 0.14 – 3. 19], PARTY BUNTING [Unit price – 2.3 – 15.79], REGENCY CAKESTAND 3 TIER [Unit price – 4 – 12.75]. These items are all reasonably priced, high discounted unit price for high quantity, which may account for their popularity. They are also all quite tiny and easy to ship, which may appeal to online customers.

7. Product-wise average quantity vs. unit price correlation: -0.25. When average unit price and average quantity are negatively associated, average quantity decreases as price rises. People buy more at a reduced price, so this makes sense. product-wise average revenue vs. average unit price correlation +0.25 Average revenue rises with average unit price when they are positively connected. It makes sense since companies who can charge more will make more money. Product-wise average revenue vs. average quantity: +0.42, Average revenue rises as average unit quantity rises when they are positively connected. Businesses that sell more items generate more money, so this makes reasonable.

6 Conclusion

Overall, the sales data shows that the business is performing well, with sales and revenue increasing over time and seasonally. This is done by providing discounts in unit price when more quantity purchased. The most popular products are those that are relatively inexpensive and small, Company more Focused on consumers on the United Kingdom as well as European countries such Germany, France, the Netherlands, EIRE, and the. The demand for goods is great in these nations, and there is a direct relationship between sales and quantity.

7 Recommendations

Noting that correlation does not imply causality is also crucial. Quantity and revenue do not necessarily cause one another, even though there is a correlation between them. It's probable that both variables are being influenced by additional factors. For instance, marketing initiatives, demand seasonality, and economic conditions can all have an impact on revenue and quantity. Make marketing and advertising investments to raise product awareness and demand. Concentrate on marketing goods that are in demand and well-liked in other countries. Customer feedback and sales data can be used to improve this.

8 References

Chen,Daqing. (2019). Online Retail II. UCI Machine Learning Repository. https://doi.org/10.24432/C5CG6D.

9 Acknowledgments

Source of the Dataset:

Dr. Daqing Chen, Course Director: MSc Data Science. chend '@' lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK.

This allows for the sharing and adaptation of the datasets for any purpose, provided that the appropriate credit is given.

Please find more information refer the below link,

https://archive.ics.uci.edu/ml/datasets/Online+Retail+II