Data Visualisation: Empowering Business with Effective Insights

Gain insights into leveraging data visualisations as a tool for making informed business decisions.

Task 1

Here is the background information on task

An online retail store has hired you as a consultant to review their data and provide insights that would be valuable to the CEO and CMO of the business. The business has been performing well and the management wants to analyse what the major contributing factors are to the revenue so they can strategically plan for next year.

The leadership is interested in viewing the metrics from both an operations and marketing perspective. Management also intends to expand the business and is interested in seeking guidance into areas that are performing well so they can keep a clear focus on what's working. They would also like to view different metrics based on the demographic information that is available in the data.

A meeting with the CEO and CMO has been scheduled for next month and you need to draft the relevant analytics and insights that would help evaluate the current business performance and suggest metrics that would enable them to make the decision on expansion.

Remember, thinking from the perspective of business leaders allows you to analyse the data more effectively and present better insights.

Here is your task

To prepare for your meeting, you need to draft questions that you think will be important and relevant to the CEO and CMO. This preparation will be your guide as you develop your presentation.

For this task, you are only required to draft the questions. Make sure to think both quantitatively and qualitatively.

You've been provided a dataset in the resources below to use as the basis for your exploration. Review this data, taking note of what information has been provided, what insights you can garner, and what is relevant to both the CEO and CMO respectively.

Create a set of four questions that you anticipate each business leader will ask and want to know the answers to. Make sure you differentiate your questions, as both the CEO and CMO view business decisions through different lenses.

<u>Questions of interest to the CEO</u>

Which region is generating the highest revenue, and which region is generating the lowest? This question is important to the CEO as it is based on the fundamental source of income for the business, i.e., revenue. Revenue analysis is important to the CEO as top-level executives are always focused on earnings and how to increase it. Here, the CEO is interested in the viewing revenue by the regions, to assess which regions are generating the highest revenue and which regions are generating lower revenue. Using the data and analysis, the CEO will be able to decide on how to further generate revenue in the regions that are already generating the most revenue. For the regions that are not generating enough revenue, the CEO will then study the reasons why there is a lack of sales in those regions and try to improve the products and make them more suitable for those regions.

What is the monthly trend of revenue, which months have faced the biggest increase/decrease?
A monthly trend of revenue will provide the CEO with insights on how the revenue is fluctuating each month. This will enable the CEO to analyze how the internal changes inside the company have had impact on the sales. E.g., how a new product launch has led to an increase in revenue during the month or how the introduction of a new region has led to an increase in revenue for the online store. The CEO can also analyze if there have been any delays internally that would have caused a potential decrease. Such analysis is vital for the senior management as it would enable them to plan ahead and try to make the customer experience as smooth as possible.

Which months generated the most revenue? Is there a seasonality in sales?
In retail businesses, there are always months that will have a greater demand due to seasonality. There will be cases where the data will experience regular and predictable changes that recur every calendar year. Such seasonal months would be necessary to identify as the CEO would be interested in devising a strategy that would gain the maximum benefit from the months that have greater demands.

Who are the top customers and how much do they contribute to the total revenue? Is the business dependent on these customers or is the customer base diversified?
This analysis is highly important as it would enable the CEO to identify what the main drivers are behind the total revenue. Looking at the top customers of the retail store would provide an idea of which customers are contributing the most to the revenue. The store can then derive a strategy where the top customers can be targeted with more products that they can buy. This will ensure higher revenue for the store as these customers
are the top buyers from the store. Although having fewer customers buying in high volumes can be beneficial for a business, there can also be a drawback. Retailers would have less bargaining power with these customers because they drive the majority of the revenue for the store and can negotiate lower prices. The CEO needs to be notified of the diversification of the customers so that he can plan ahead of time. In cases wehre the business is highly dependent on a few customers, the plan would be to increase the customer base and target more customers that would bring more revenue to the store.

<u>Questions of interest to the CMO</u>

What is the percentage of customers who are repeating their orders? Are they ordering the same products or different?
This question shows that the CMO is interested in viewing the trends in customer orders. He is interested to know how many customers out of the total are coming back to them and re-ordering. This analysis will help explain to the CMO what percentage of customers are buying from them more than once. Once this is identified, the CMO can come up with a strategy to target these customers with more offers and products that they would need. The analysis will also be done to see what they are buying the second time, this will provide the CMO trends into what products and sub products are in demand and then a marketing strategy can be devised to target these customers with better options.

For the repeat customers, how long does it take for them to place the next order after being delivered the previous one?
This analysis will help the CMO identify the frequency of orders. This would mean determining how long the customers are taking to re-order from the store. The expectation is that those customers who have recently made a purchase would have the product on their mind and are expected to purchase or use the product again in the future. Once the information is gathered from the analysis, the CMO can create a strategy to get the recent customers to revisit the business and spend more. For the customers who have not made purchases again from

the store, efforts can be made to remind them that it has been a while since they last purchased from the store. Incentivizing customers also comes into play in this scenario.

What revenue is being generated from the customers who have ordered more than once? Revenue stems from how much the customer spends to purchase the products from the store. Therefore, the analysis needs to be done to determine how much revenue is being generated from the customers who are regular buyers from the store. The CMO can devise a strategy to encourage customers who spend more money on repeat purchases to continue to do so. It is also important to note that if a customer has made a big purchase the first time, they should be encouraged to come and shop from the store again. A marketing strategy will ensure that the high paying customers will continue to bring more revenue to the store going forward, as well.

Who are the customers that have repeated the most? How much are they contributing to revenue? It is also important to assess which customers are repeating the most and how much are they contributing to the revenue. There would be customers who need the same products on a weekly or monthly basis, however, the products do not have a high monetary value. Therefore, the contribution to revenue for these customers will be low. On the other hand, there might be customers who are ordering twice a year and have very big orders in terms of revenue. These customers buy on certain months only, therefore, the management needs to ensure that enough supplies are available to accommodate their orders. The customers will high order volumes and low revenue would need to be offered more discounts so that they can buy in bulk and lead to more revenue.

Task 2

Here is the background information on task

You have been asked by the CEO and CMO to provide visuals on the metrics that they wish to analyse for the online retail store. You will gather the requirements and provide them with the type of visual that would be best suited to the scenario. The senior management wants to understand how their business is performing and what areas are the key strengths of the company. They are also focused on identifying opportunities that would lead to growth and generate more revenue in the future.

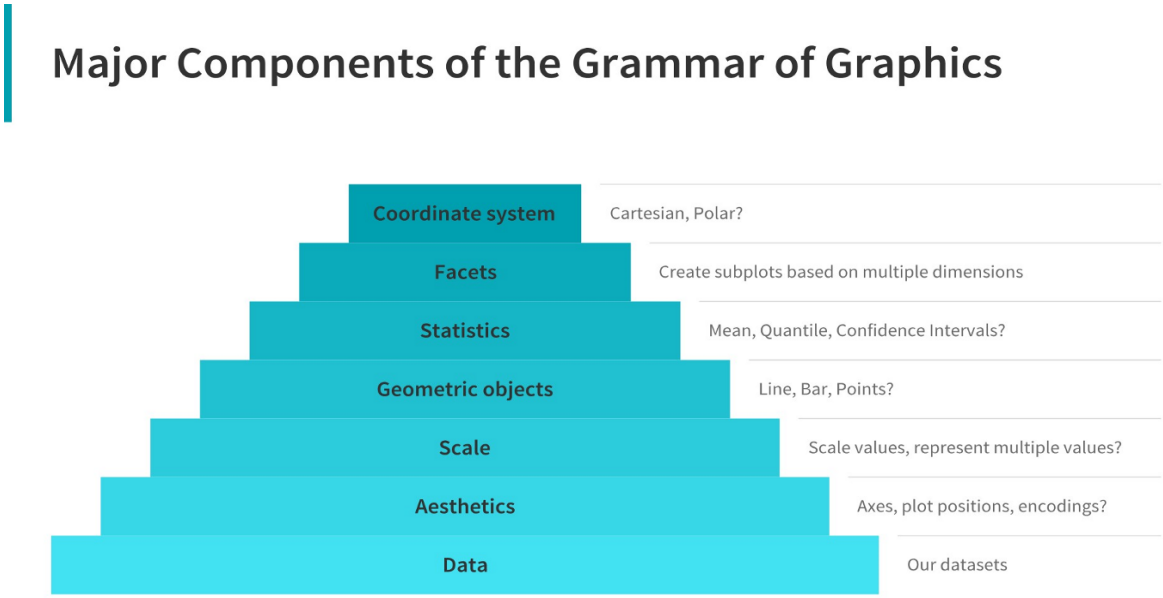
You will be provided multiple visualisation requests by the CEO and you would need to provide the visual which would explain the data and insights in the simplest possible manner. The visual should adequately convey the information that you are trying to present. This exercise is critical to the senior management as any incorrect representation can lead to a wrong message being conveyed, or a wrong decision taken by the management. Therefore, you would need to make sure that the correct visuals are used to represent each set of data.

Here is task

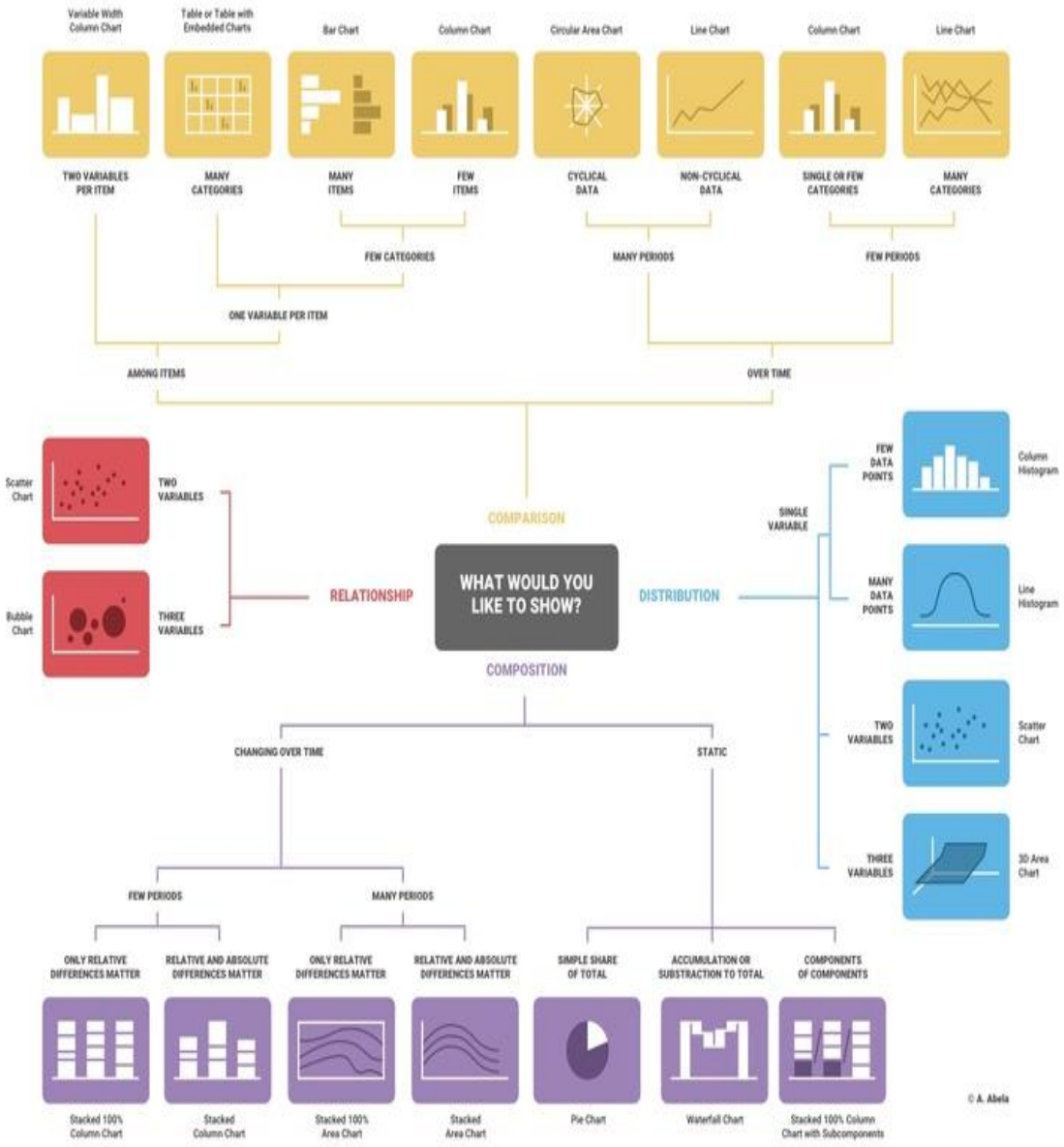
In this task, you will be required to read the questions carefully and understand that business requirement. Once you have an idea of what is required from the perspective of the CEO and CMO, you will need to come up with the perfect visual which will illustrate what the senior managers are looking for in each scenario. Remember, data can be presented in multiple types of charts, but you are required to select the visual that would best display the information which is being presented.

You will be provided resources on how to select visuals based on the different scenarios, these are available in the resources below. These resources will help you get an idea on which visual to select for the given business scenario and will also guide you on how to choose the right chart or graph for your data. Each question will contain a unique scenario and you will be expected to answer the questions based on that scenario.

Grammar of Graphics to understand the dimensions of visuals.



Choosing right charts based upon need such as Comparison, Distribution, Relationship, and Composition.



Task 3

Here is the background information on task

It's time to present your findings to the CEO and CMO. They are interested in your thought process and how you have handled the data cleanup and visualisation phase. It is important to explain your thought process and ideas in a clear and straightforward way. You are also required to clearly present the analysis of all four questions from the previous task. Make sure you are well versed with the data and the conclusions you've made from your analysis.

Expansion is top of mind for these leaders and they're keen to understand where the most lucrative opportunities are in their business**.**

Here is task

Develop a script and record a video presenting your findings to the CEO and CMO based on the four questions they asked and the visuals you created in the previous tasks.

You can use your work or the model answer from the previous task to develop your presentation.

When writing your script, you should speak about your entire process, including the initial data load and clean-up steps so that your leaders know you've done your due diligence in providing error-free analysis. Data analysis provides heaps of information but remember to focus on the information that is most important to your leaders.

Sample Answer

Good Afternoon,

I'm Deepankaj and I'm excited to share some insights about your business. Thank you for providing the guiding questions. It was helpful to see what types of insights you are looking to gain from the data. I hope you find the analysis compelling and helpful as you make decisions regarding future business opportunities.

First off, I want to assure you that I've provided the most up to date and error free analysis using python. After I loaded the data into my software, I scrubbed any records that have negative quantities and unit price, as these records needed to be removed in order to provide helpful analysis.

As for your first question, the CEO has requested a trend of the revenue to see if there is any seasonality in the store sales. My analysis shows that there are some months of the year where exceptional growth is witnessed. The data shows that the revenue in the first 8 months is fairly constant as the average revenue generated for these 8 months is around $685k. The increase in revenue starts in the month of September, where the revenue increases by 40% over the previous month. This trend continues till the month of November where it reached 1.5 million USD, the highest during the entire year. The data is incomplete for the month of December, therefore, no conclusion can be drawn from it, unfortunately. This analysis shows that the retail store sales are impacted by the seasonality which usually occurs in the last 4 months of the year.

The second visual shows how the top 10 countries which have opportunities for growth are performing. This data does not include the UK as the country already has high demand and I've been told you're more focused on the countries where demand can be increased. The analysis shows that countries such as the Netherlands, Ireland, Germany and France have high volumes of units bought and revenue generated. I would suggest that these countries should be focused on to ensure that measures are taken to capture these markets even more.

The third analysis has been performed on the top 10 customers who have purchased the most from the store. The data shows that there is not much of a difference between the purchases made by the top 10 customers. The highest revenue generating customer only purchased 17% more than the 2nd highest which shows that the business is not relying only on a few customers to generate the revenue. This shows that the bargaining power of customers is low and the business is in a good position.

Finally, the map chart shows the regions that have generated the most revenue compared with the regions that have not. It can be seen that apart from the UK, countries such as Netherlands, Ireland, Germany, France and Australia are generating high revenue and the company should invest more in these areas to increase demand for products. The map also shows that most of the sales are only in the European region with very few in the American region. Africa and Asia do not have any demand for the products, along with Russia. A new strategy targeting these areas has the potential to boost sales revenues and profitability.

Thanks so much for your time. If you have any questions about the analysis or would like to see anything additional after you've had time to digest this information, I'd be happy to develop that for you.

In [18]:
```
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
```

In [20]:
```
df=pd.read_excel(r"C:\Users\deepa\OneDrive\Desktop\Online_Retail_Data_Set.xlsx")
```

In [21]:
```
df.head()
```

Out[21]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

In [46]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   InvoiceNo    541909 non-null  object 
 1   StockCode    541909 non-null  object 
 2   Description  540455 non-null  object 
 3   Quantity     541909 non-null  int64  
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object 
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
```

In [55]:
```
df.isna().sum()
```

Out[55]:

```
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID    135080
Country            0
dtype: int64
```

In [56]:
```
df_neg_qmdf[df["Quantity"]<1]
```

In [77]:
```
df_pos_qmdf[df["Quantity"]>0]
```

In [81]:
```
df_pos_q.shape
```

Out[81]:
```
(531285, 8)
```

In [59]:
```
df_neg_q.shape
```

Out[59]:
```
(10624, 8)
```

In [109]:
```
df_neg_pmdf_pos_q[df["UnitPrice"]<0]
```

```
C:\Users\deepa\AppData\Local\Temp\ipykernel_14880\2430195404.py:1: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  df_neg_p=df_pos_q[df["UnitPrice"]<0]
```

In [111]:
```
df_neg_p.shape
```

Out[111]:
```
(2, 8)
```

In [122]:
```
df_pos_pmdf_pos_q[df_pos_q["UnitPrice"]>0]
df_pos_p.shape
```

Out[122]:
```
(530104, 8)
```

In [133]:
```
df=df_pos_p.copy()
```

In [146]:
```
df.shape
```

Out[146]:
```
(530104, 8)
```

In [155]:
```
for i,j in zip(df.columns,df.isna().sum()):
    print(i,j,"-",round(j/df.shape[0],4)*100,"%")
```

```
InvoiceNo 0 - 0.0 %
StockCode 0 - 0.0 %
Description 0 - 0.0 %
Quantity 0 - 0.0 %
InvoiceDate 0 - 0.0 %
UnitPrice 0 - 0.0 %
CustomerID 132220 - 24.94 %
Country 0 - 0.0 %
```

In [166]:
```
# Missing value imputation

df["Description"].fillna(df["Description"].mode()[0],inplace=True)
```

```
C:\Users\deepa\AppData\Local\Temp\ipykernel_14880\2867247727.py:3: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.


  df["Description"].fillna(df["Description"].mode()[0],inplace=True)
```

In [157]:
```
for i,j in zip(df.columns,df.isna().sum()):
    print(i,j,"-",round(j/df.shape[0],4)*100,"%")
```

```
InvoiceNo 0 - 0.0 %
StockCode 0 - 0.0 %
Description 0 - 0.0 %
Quantity 0 - 0.0 %
InvoiceDate 0 - 0.0 %
UnitPrice 0 - 0.0 %
CustomerID 132220 - 24.94 %
Country 0 - 0.0 %
```

In [158]:
```
df["CustomerID"].nunique()
```

Out[158]:
```
4338
```

In [159]:
```
# Missing value imputation by bfill or ffill or mode in Customer ID column

df["CustomerID"]=df["CustomerID"].fillna(method="ffill")

#  df["CustomerID"]=df["CustomerID"].fillna(method="bfill")
```

```
C:\Users\deepa\AppData\Local\Temp\ipykernel_14880\2283629695.py:3: FutureWarning: Series.fillna with 'method' is deprecated and will raise in a future version. Use obj.ffill() or obj.bfill() instead.
  df["CustomerID"]=df["CustomerID"].fillna(method="ffill")
```

In [220]:
```
df["CustomerID"].isna().sum()
```

Out[220]:
```
0
```

In [221]:
```
for i,j in zip(df.columns,df.isna().sum()):
    print(i,j,"-",round(j/df.shape[0],4)*100,"%")
```

```
InvoiceNo 0 - 0.0 %
StockCode 0 - 0.0 %
Description 0 - 0.0 %
Quantity 0 - 0.0 %
InvoiceDate 0 - 0.0 %
UnitPrice 0 - 0.0 %
CustomerID 0 - 0.0 %
Country 0 - 0.0 %
```

In [222]:
```
# Basic summary statistics

df.describe()
```

Out[222]:

	Quantity	InvoiceDate	UnitPrice	CustomerID
count	530104.000000	530104	530104.000000	530104.000000
mean	10.542037	2011-07-04 20:16:05.225087744	3.907625	15287.610047
min	1.000000	2010-12-01 08:26:00	0.001000	12346.000000
25%	1.000000	2011-03-28 12:22:00	1.250000	13804.000000
50%	3.000000	2011-07-20 12:58:00	2.080000	15179.000000
75%	10.000000	2011-10-19 12:39:00	4.130000	16813.000000
max	80995.000000	2011-12-09 12:50:00	13541.330000	18287.000000
std	155.524124	NaN	35.915681	1735.660857

In [223]:
```
# creating new columns as total_amt

df["total_amt"]=df["Quantity"]*df["UnitPrice"]
```

In [224]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 530104 entries, 0 to 541908
Data columns (total 9 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   InvoiceNo    530104 non-null  object 
 1   StockCode    530104 non-null  object 
 2   Description  530104 non-null  object 
 3   Quantity     530104 non-null  int64  
 4   InvoiceDate  530104 non-null  datetime64[ns]
 5   UnitPrice    530104 non-null  float64
 6   CustomerID   530104 non-null  float64
 7   Country      530104 non-null  object 
 8   total_amt    530104 non-null  float64
dtypes: datetime64[ns](1), float64(3), int64(1), object(4)
memory usage: 40.4+ MB
```

In [225]:
```
df.describe()
```

Out[225]:

	Quantity	InvoiceDate	UnitPrice	CustomerID	total_amt
count	530104.000000	530104	530104.000000	530104.000000	530104.000000
mean	10.542037	2011-07-04 20:16:05.225087744	3.907625	15287.610047	20.121871
min	1.000000	2010-12-01 08:26:00	0.001000	12346.000000	0.001000
25%	1.000000	2011-03-28 12:22:00	1.250000	13804.000000	3.750000
50%	3.000000	2011-07-20 12:58:00	2.080000	15179.000000	9.900000
75%	10.000000	2011-10-19 12:39:00	4.130000	16813.000000	17.700000
max	80995.000000	2011-12-09 12:50:00	13541.330000	18287.000000	168469.600000
std	155.524124	NaN	35.915681	1735.660857	270.356743

In [226]:
```
df["InvoiceDate"]=df["InvoiceDate"].astype("str")
```

In [227]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 530104 entries, 0 to 541908
Data columns (total 9 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   InvoiceNo    530104 non-null  object 
 1   StockCode    530104 non-null  object 
 2   Description  530104 non-null  object 
 3   Quantity     530104 non-null  int64  
 4   InvoiceDate  530104 non-null  object 
 5   UnitPrice    530104 non-null  float64
 6   CustomerID   530104 non-null  float64
 7   Country      530104 non-null  object 
 8   total_amt    530104 non-null  float64
dtypes: float64(3), int64(1), object(5)
memory usage: 40.4+ MB
```

In [228]:
```
df["Date"]=df["InvoiceDate"].str.split(" ")
```

In [229]:
```
#df.drop(columns="Year",axis=1,inplace=True)
```

In [230]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 530104 entries, 0 to 541908
Data columns (total 10 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   InvoiceNo    530104 non-null  object 
 1   StockCode    530104 non-null  object 
 2   Description  530104 non-null  object 
 3   Quantity     530104 non-null  int64  
 4   InvoiceDate  530104 non-null  object 
 5   UnitPrice    530104 non-null  float64
 6   CustomerID   530104 non-null  float64
 7   Country      530104 non-null  object 
 8   total_amt    530104 non-null  float64
 9   Date         530104 non-null  object 
dtypes: float64(3), int64(1), object(6)
memory usage: 44.5+ MB
```

In [231]:
```
df["Date"]=df["Date"][0][0]
```

In [232]:
```
df["Year"]=df["Date"].str.split("-")[0][0]
```

In [233]:
```
df["Month"]=df["Date"].str.split("-")[0][1]
```

In [234]:
```
df["Year"]=df["Year"].astype("int")
df["Month"]=df["Month"].astype("int")
```

In [235]:
```
df["Year"].isna().sum()
```

Out[235]:
```
0
```

In [261]:
```
df["Month"].isna().sum()
```

Out[261]:
```
0
```

In [237]:
```
df["Date"]=pd.to_datetime(df["Date"])
```

In [238]:
```
df.describe(include="O")
```

Out[238]:

	InvoiceNo	StockCode	Description	InvoiceDate	Country
count	530104	530104	530104	530104	530104
unique	19960	3922	4026	18499	38
top	573585	85123A	WHITE HANGING HEART T-LIGHT HOLDER	2011-10-31 14:41:00	United Kingdom
freq	1114	2265	2323	1114	485123

In [239]:
```
df.describe()
```

Out[239]:

	Quantity	UnitPrice	CustomerID	total_amt	Date	Year	Month
count	530104.000000	530104.000000	530104.000000	530104.000000	530104	530104.0	530104.0
mean	10.542037	3.907625	15287.610047	20.121871	2010-12-01 00:00:00	2010.0	12.0
min	1.000000	0.001000	12346.000000	0.001000	2010-12-01 00:00:00	2010.0	12.0
25%	1.000000	1.250000	13804.000000	3.750000	2010-12-01 00:00:00	2010.0	12.0
50%	3.000000	2.080000	15179.000000	9.900000	2010-12-01 00:00:00	2010.0	12.0
75%	10.000000	4.130000	16813.000000	17.700000	2010-12-01 00:00:00	2010.0	12.0
max	80995.000000	13541.330000	18287.000000	168469.600000	2010-12-01 00:00:00	2010.0	12.0
std	155.524124	35.915681	1735.660857	270.356743	NaN	0.0	0.0

In [240]:
```
df.to_csv(r"C:\Users\deepa\OneDrive\Desktop\Online_Retail_Data_Set1.csv",index=False)
```

In [ ]:
```

```

In [ ]:
```

```