Company Information:

Bicree is a 3PL AI tech enabled logistic aggregator platform, which provides customers with quality automated shipping solutions. Its services offer opportunities to ship anywhere in India. Bicree's AI-powered dashboard is highly compatible with most e-commerce platforms and marketplaces. It is also integrable with custom websites and offers rest APIs. Bicree's leading and specialized shipping partners ensure a safe and fast order delivery across varying volumes, type or size. It provides shipping across multiple segments ranging from apparels to electronics. Their platform helps in monitoring orders easily and increase profitability. It allows customers to analyze, connect, shipping assistance, auto NDR, COD Verification and Geo Analysis.

In [ ]:

Task : Data Cleaning

Data cleaning, or data wrangling or data remediation, or data munging, refers to a variety of processes designed to transform raw data into more readily used formats. It is the method of cleaning, arranging, and enriching raw information into the desired format for better decision-making in less time. The exact methods differ from project to project depending on the data you're leveraging and the goal you're trying to achieve.

Importing required libraries

In [105…

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
plt.rcParams['figure.figsize']=(18,7)
```

Importing Dataset

In [106…

```
df=pd.read_csv(r"C:\Users\rakes\Downloads\Bicree_shipment- test - SCMS_Delivery_Hist
pd.set_option('display.max_columns', None)
df.head(25)
```

Out[106…

	ID	Project Code	PQ #	PO / SO #	ASN/DN #	Country	Managed By	Fulfill Via	Vendor INCO Term	Shipment Mode	PC Firs Sen to Clien Date
0	1	100-CI-T01	Pre-PQ Process	SCMS-4	ASN-8	Côte d'Ivoire	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces
1	3	108-VN-T01	Pre-PQ Process	SCMS-13	ASN-85	Vietnam	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces

	ID	Project Code	PQ #	PO / SO #	ASN/DN #	Country	Managed By	Fulfill Via	Vendor INCO Term	Shipment Mode	PC Firs Sen to Clien Date
2	4	100-CI-T01	Pre-PQ Process	SCMS-20	ASN-14	Côte d'Ivoire	PMO - US	Direct Drop	FCA	Air	Pre-PC Proces
3	15	108-VN-T01	Pre-PQ Process	SCMS-78	ASN-50	Vietnam	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces
4	16	108-VN-T01	Pre-PQ Process	SCMS-81	ASN-55	Vietnam	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces
5	23	112-NG-T01	Pre-PQ Process	SCMS-87	ASN-57	Nigeria	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces
6	44	110-ZM-T01	Pre-PQ Process	SCMS-139	ASN-130	Zambia	PMO - US	Direct Drop	DDU	Air	Pre-PC Proces
7	45	109-TZ-T01	Pre-PQ Process	SCMS-140	ASN-94	Tanzania	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces
8	46	112-NG-T01	Pre-PQ Process	SCMS-156	ASN-93	Nigeria	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces
9	47	110-ZM-T01	Pre-PQ Process	SCMS-165	ASN-199	Zambia	PMO - US	Direct Drop	CIP	Air	Pre-PC Proces
10	60	110-ZM-T01	Pre-PQ Process	SCMS-221	ASN-223	Zambia	PMO - US	Direct Drop	CIP	Air	Pre-PC Proces
11	61	110-ZM-T01	Pre-PQ Process	SCMS-226	ASN-137	Zambia	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces
12	62	102-NG-T01	Pre-PQ Process	SCMS-230	ASN-144	Nigeria	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces
13	64	107-RW-T01	Pre-PQ Process	SCMS-268	ASN-242	Rwanda	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces
14	65	106-HT-T01	Pre-PQ Process	SCMS-274	ASN-162	Haiti	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces
15	68	113-ZW-T01	Pre-PQ Process	SCMS-308	ASN-285	Zimbabwe	PMO - US	Direct Drop	CIP	Air	Pre-PC Proces

	ID	Project Code	PQ #	PO / SO #	ASN/DN #	Country	Managed By	Fulfill Via	Vendor INCO Term	Shipment Mode	PC Firs Sen to Clien Date
16	69	102-NG-T01	Pre-PQ Process	SCMS-354	ASN-608	Nigeria	PMO - US	Direct Drop	CIP	NaN	Pre-PC Proces
17	80	107-RW-T01	Pre-PQ Process	SCMS-488	ASN-299	Rwanda	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces
18	87	109-TZ-T01	Pre-PQ Process	SCMS-555	ASN-409	Tanzania	PMO - US	Direct Drop	CIP	Air	Pre-PC Proces
19	92	102-NG-T01	Pre-PQ Process	SCMS-592	ASN-485	Nigeria	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces
20	96	102-NG-T01	Pre-PQ Process	SCMS-570	ASN-451	Nigeria	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces
21	108	104-CI-T01	Pre-PQ Process	SCMS-698	ASN-727	Côte d'Ivoire	PMO - US	Direct Drop	CIP	Air	Pre-PC Proces
22	115	108-VN-T01	Pre-PQ Process	SCMS-753	ASN-781	Vietnam	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces
23	116	108-VN-T01	Pre-PQ Process	SCMS-759	ASN-632	Vietnam	PMO - US	Direct Drop	FCA	Air	Pre-PC Proces
24	130	100-HT-T01	Pre-PQ Process	SCMS-10080	ASN-628	Haiti	PMO - US	Direct Drop	EXW	Air	Pre-PC Proces

Checking the number of rows and columns

In [107…

```
df.shape
print('In this provided data we have {} number of Rows and {} number of columns'.for
```

```
In this provided data we have 10324 number of Rows and 33 number of columns
```

In [108…

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10324 entries, 0 to 10323
Data columns (total 33 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   ID                             10324 non-null  int64
 1   Project Code                   10324 non-null  object
 2   PQ #                           10324 non-null  object
 3   PO / SO #                      10324 non-null  object
```

```
 4   ASN/DN #                     10324 non-null  object
 5   Country                      10324 non-null  object
 6   Managed By                   10324 non-null  object
 7   Fulfill Via                  10324 non-null  object
 8   Vendor INCO Term             10324 non-null  object
 9   Shipment Mode                9964 non-null   object
 10  PQ First Sent to Client Date  10324 non-null  object
 11  PO Sent to Vendor Date       10324 non-null  object
 12  Scheduled Delivery Date      10324 non-null  object
 13  Delivered to Client Date     10324 non-null  object
 14  Delivery Recorded Date       10324 non-null  object
 15  Product Group                10324 non-null  object
 16  Sub Classification           10324 non-null  object
 17  Vendor                       10324 non-null  object
 18  Item Description             10324 non-null  object
 19  Molecule/Test Type           10324 non-null  object
 20  Brand                        10324 non-null  object
 21  Dosage                       8588 non-null   object
 22  Dosage Form                  10324 non-null  object
 23  Unit of Measure (Per Pack)   10324 non-null  int64
 24  Line Item Quantity           10324 non-null  int64
 25  Line Item Value              10324 non-null  float64
 26  Pack Price                   10324 non-null  float64
 27  Unit Price                   10324 non-null  float64
 28  Manufacturing Site           10324 non-null  object
 29  First Line Designation       10324 non-null  object
 30  Weight (Kilograms)           10324 non-null  object
 31  Freight Cost (USD)           10324 non-null  object
 32  Line Item Insurance (USD)    10037 non-null  float64
dtypes: float64(4), int64(3), object(26)
memory usage: 2.6+ MB
```

Datatype of the features

In [109…

```
df.dtypes
```

Out[109…

```
ID                              int64
Project Code                   object
PQ #                           object
PO / SO #                      object
ASN/DN #                       object
Country                        object
Managed By                     object
Fulfill Via                    object
Vendor INCO Term               object
Shipment Mode                  object
PQ First Sent to Client Date   object
PO Sent to Vendor Date         object
Scheduled Delivery Date        object
Delivered to Client Date       object
Delivery Recorded Date         object
Product Group                  object
Sub Classification             object
Vendor                         object
Item Description               object
Molecule/Test Type             object
Brand                          object
Dosage                         object
Dosage Form                    object
Unit of Measure (Per Pack)      int64
Line Item Quantity              int64
Line Item Value                float64
Pack Price                     float64
Unit Price                     float64
Manufacturing Site              object
First Line Designation          object
Weight (Kilograms)              object
```

```
Freight Cost (USD)              object
Line Item Insurance (USD)      float64
dtype: object
```

Columns of the data

In [110…

```
columns=df.columns
columns
```

Out[110…

```
Index(['ID', 'Project Code', 'PQ #', 'PO / SO #', 'ASN/DN #', 'Country',
       'Managed By', 'Fulfill Via', 'Vendor INCO Term', 'Shipment Mode',
       'PQ First Sent to Client Date', 'PO Sent to Vendor Date',
       'Scheduled Delivery Date', 'Delivered to Client Date',
       'Delivery Recorded Date', 'Product Group', 'Sub Classification',
       'Vendor', 'Item Description', 'Molecule/Test Type', 'Brand', 'Dosage',
       'Dosage Form', 'Unit of Measure (Per Pack)', 'Line Item Quantity',
       'Line Item Value', 'Pack Price', 'Unit Price', 'Manufacturing Site',
       'First Line Designation', 'Weight (Kilograms)', 'Freight Cost (USD)',
       'Line Item Insurance (USD)'],
      dtype='object')
```

Categorical features

In [111…

```
cat_feat=df.select_dtypes(include='object')
print("The categorical Features of the data are:",'\n',cat_feat.columns)
```

```
The categorical Features of the data are:
 Index(['Project Code', 'PQ #', 'PO / SO #', 'ASN/DN #', 'Country',
       'Managed By', 'Fulfill Via', 'Vendor INCO Term', 'Shipment Mode',
       'PQ First Sent to Client Date', 'PO Sent to Vendor Date',
       'Scheduled Delivery Date', 'Delivered to Client Date',
       'Delivery Recorded Date', 'Product Group', 'Sub Classification',
       'Vendor', 'Item Description', 'Molecule/Test Type', 'Brand', 'Dosage',
       'Dosage Form', 'Manufacturing Site', 'First Line Designation',
       'Weight (Kilograms)', 'Freight Cost (USD)'],
      dtype='object')
```

Frequency of the categorical features

In [112…

```
for i in cat_feat.columns:
    print(i,'\n',cat_feat[i].value_counts(),'\n')
```

```
Project Code
 116-ZA-T30    768
104-CI-T30     729
151-NG-T30     628
114-UG-T30     596
108-VN-T30     522
              ...
900-GY-T30       1
103-UG-T30       1
100-SL-T01       1
201-UG-T30       1
A01-SN-T50       1
Name: Project Code, Length: 142, dtype: int64

PQ #
 Pre-PQ Process    2681
FPQ-14942           205
FPQ-12522           154
FPQ-13973           110
FPQ-4537             98
                   ...
FPQ-5295              1
```

```
FPQ-10107            1
FPQ-11585            1
FPQ-4106             1
FPQ-11773            1
Name: PQ #, Length: 1237, dtype: int64

PO / SO #
 SCMS-199289    67
SCMS-199283    63
SCMS-183950    55
SCMS-259075    38
SCMS-215370    38
               ..
SCMS-33930      1
SCMS-116850     1
SCMS-181381     1
SO-44141        1
SCMS-68330      1
Name: PO / SO #, Length: 6233, dtype: int64

ASN/DN #
 ASN-19166    54
ASN-24415    38
ASN-23875    26
ASN-32138    19
DN-304       17
             ..
DN-2286       1
DN-205        1
ASN-25268     1
ASN-23076     1
ASN-2569      1
Name: ASN/DN #, Length: 7030, dtype: int64

Country
 South Africa          1406
Nigeria               1194
Côte d'Ivoire         1083
Uganda                 779
Vietnam                688
Zambia                 683
Haiti                  655
Mozambique             631
Zimbabwe               538
Tanzania               519
Rwanda                 430
Congo, DRC             333
Guyana                 237
Ethiopia               216
South Sudan            164
Kenya                  111
Burundi                 98
Namibia                 95
Cameroon                75
Botswana                70
Ghana                   58
Dominican Republic      52
Sudan                   46
Swaziland               35
Mali                    17
Guatemala               15
Pakistan                15
Malawi                  14
Benin                   13
Lebanon                  8
Libya                    8
Angola                   7
Liberia                  6
Sierra Leone             4
```

```
Lesotho                  4
Togo                     3
Afghanistan              3
Senegal                  3
Kazakhstan               2
Kyrgyzstan               2
Burkina Faso             2
Guinea                   1
Belize                   1
Name: Country, dtype: int64

Managed By
 PMO - US                      10265
South Africa Field Office         57
Ethiopia Field Office              1
Haiti Field Office                 1
Name: Managed By, dtype: int64

Fulfill Via
 From RDC       5404
Direct Drop    4920
Name: Fulfill Via, dtype: int64

Vendor INCO Term
 N/A - From RDC    5404
EXW               2778
DDP               1443
FCA                397
CIP                275
DDU                 15
DAP                  9
CIF                  3
Name: Vendor INCO Term, dtype: int64

Shipment Mode
 Air            6113
Truck          2830
Air Charter     650
Ocean           371
Name: Shipment Mode, dtype: int64

PQ First Sent to Client Date
 Pre-PQ Process       2476
Date Not Captured     205
9/11/14               205
7/11/13               173
4/30/14               123
                      ...
11/4/09                 1
8/27/14                 1
12/23/13                1
3/22/11                 1
1/4/12                  1
Name: PQ First Sent to Client Date, Length: 765, dtype: int64

PO Sent to Vendor Date
 N/A - From RDC       5404
Date Not Captured     328
8/27/14                80
3/19/10                78
8/29/14                76
                      ...
5/1/12                  1
7/13/09                 1
10/22/09                1
7/13/11                 1
3/1/07                  1
Name: PO Sent to Vendor Date, Length: 897, dtype: int64
```

```
Scheduled Delivery Date
 29-Aug-14    97
16-Mar-12    83
27-Aug-14    63
31-May-10    62
31-Jan-14    60
             ..
3-Jul-13      1
12-May-07     1
3-Oct-07      1
2-Jan-13      1
8-Aug-12      1
Name: Scheduled Delivery Date, Length: 2006, dtype: int64

Delivered to Client Date
 29-Aug-14    74
27-Aug-14    66
28-Jun-10    60
14-Feb-12    60
16-Apr-13    59
             ..
16-May-08     1
3-Nov-10      1
30-Sep-06     1
21-Jul-08     1
20-May-13     1
Name: Delivered to Client Date, Length: 2093, dtype: int64

Delivery Recorded Date
 29-Aug-14    67
27-Aug-14    66
28-Jun-10    60
14-Feb-12    60
16-Apr-13    59
             ..
5-May-10      1
28-Jul-10     1
28-Aug-12     1
28-Sep-07     1
30-Aug-10     1
Name: Delivery Recorded Date, Length: 2042, dtype: int64

Product Group
 ARV     8550
HRDT    1728
ANTM      22
ACT       16
MRDT       8
Name: Product Group, dtype: int64

Sub Classification
 Adult                   6595
Pediatric               1955
HIV test                1567
HIV test - Ancillary     161
Malaria                   30
ACT                       16
Name: Sub Classification, dtype: int64

Vendor
 SCMS from RDC               5404
Orgenics, Ltd                754
S. BUYS WHOLESALER           715
Aurobindo Pharma Limited     668
Trinity Biotech, Plc         356
                            ...
MEDMIRA EAST AFRICA LTD.       1
OMEGA DIAGNOSTICS LTD          1
ACTION MEDEOR E.V.             1
```

```
ACCESS BIO, INC.              1
SETEMA LIMITED PLC            1
Name: Vendor, Length: 73, dtype: int64

Item Description
 Efavirenz 600mg, tablets, 30 Tabs                                          75
5
Nevirapine 200mg, tablets, 60 Tabs                                         623
Lamivudine/Zidovudine 150/300mg, tablets, 60 Tabs                          597
Lamivudine/Nevirapine/Zidovudine 150/200/300mg, tablets, 60 Tabs           580
HIV 1/2, Determine Complete HIV Kit, 100 Tests                             577
                                                                           ...
Lopinavir/Ritonavir 80/20mg/ml [Kaletra], oral solution, cool, Bottle, 160 ml      1
Nevirapine 50mg, dispersible tablets, 60 Tabs                                      1
Lamivudine/Zidovudine+Abacavir 150/300+300mg, tablets, co-blister, 60+60 Tabs      1
HIV 1/2, Visitect Kit, 25 Tests                                                    1
Malaria, Antigen P.f., HRP2 CareStart Kit, 60 Tests                                1
Name: Item Description, Length: 184, dtype: int64

Molecule/Test Type
 Efavirenz
1125
Nevirapine
877
Lamivudine/Nevirapine/Zidovudine
707
Lamivudine/Zidovudine
689
Lopinavir/Ritonavir
633

...
Primaquine base (as diphosphate)
1
HIV 1/2, Colloidal Gold, Diagnostic Kit Set (includes lancet, transfer pipette & alc
ohol prep pad)       1
HIV 1/2, INSTI HIV 1/2 Antibody Kit
1
Malaria, Antigen P.f., HRP2 CareStart Kit
1
Malaria Antigen P.f , HRP2, Kit
1
Name: Molecule/Test Type, Length: 86, dtype: int64

Brand
 Generic            7285
Determine           799
Uni-Gold            373
Aluvia              250
Kaletra             165
Norvir              136
Stat-Pak            115
Bioline             113
Truvada              94
Videx                84
Colloidal Gold       70
Stocrin/Sustiva      69
OraQuick             60
Invirase             53
Viread               52
Zerit                46
Isentress            44
Epivir               42
Prezista             42
Retrovir             41
Videx EC             41
Ziagen               37
Crixivan             36
Capillus             35
```

```
Intelence            32
Genie                30
Viramune             28
Clearview            19
Reyataz              18
Trizivir             18
Atripla              16
First Response       15
Coartem              12
Viracept             11
INSTi                 5
DoubleCheck           5
Paramax               5
Multispot             5
LAV                   4
Combivir              3
Hexagon               3
ImmunoComb            3
Reveal                3
InstantCHEK           2
Bundi                 2
Visitect              1
Pepti-LAV             1
CareStart             1
Name: Brand, dtype: int64

Dosage
 300mg              990
200mg               932
600mg               772
150/300mg           600
150/300/200mg       580
10mg/ml             552
150mg               431
200/50mg            395
300/300mg           301
600/300/300mg       286
150/200/30mg        250
100mg               228
50mg                174
200/300mg           160
80/20mg/ml          158
400mg               156
20mg/ml             152
30mg                144
600/200/300mg       139
150/30mg            133
30/50/60mg          127
300/200mg            94
30/60mg              89
250mg                88
60/30mg              73
100/25mg             73
600/300mg            63
300/100mg            54
1mg/ml               54
20mg                 43
25mg                 39
15mg                 38
30mg/ml              33
300/150/300mg        28
30/50/6mg            19
30/6mg               14
500/25mg             13
80mg/ml              13
60/100/12mg          12
20/120mg             12
2g                   11
500mg                10
```

```
150/300+200mg        8
133.3/33.3mg         7
150/300mg+600mg      7
40mg                 6
150/200/40mg         5
60/12mg              5
60mg                 5
50+153mg             4
125mg                4
600mg/2ml            2
50mg/g               1
150/300+300mg        1
Name: Dosage, dtype: int64

Dosage Form
 Tablet                              3532
Tablet - FDC                        2749
Test kit                            1575
Capsule                              729
Oral solution                        727
Chewable/dispersible tablet - FDC    239
Oral suspension                      214
Test kit - Ancillary                 161
Chewable/dispersible tablet          146
Delayed-release capsules             131
Delayed-release capsules - blister    41
Powder for oral solution              28
Tablet - FDC + co-blister             20
Tablet - FDC + blister                15
Tablet - blister                      10
Injection                              6
Oral powder                            1
Name: Dosage Form, dtype: int64

Manufacturing Site
 Aurobindo Unit III, India                                                3172
Mylan (formerly Matrix) Nashik                                           1415
Hetero Unit III Hyderabad IN                                              869
Cipla, Goa, India                                                         665
Strides, Bangalore, India.                                                540
                                                                         ...
ABBVIE Labs North Chicago US                                                1
Weifa A.S., Hausmanngt. 6, P.O. Box 9113 GrÃ¸nland, 0133, Oslo, Norway       1
BUNDI INTERNATIONAL DIAGNOSTICS LTD                                         1
GSK Barnard Castle UK                                                       1
Meditab (for Cipla) Daman IN                                                1
Name: Manufacturing Site, Length: 88, dtype: int64

First Line Designation
 Yes    7030
No     3294
Name: First Line Designation, dtype: int64

Weight (Kilograms)
 Weight Captured Separately    1507
2                               29
6                               26
1                               23
60                              20
                               ...
1281                             1
11283                            1
2842                             1
2515                             1
1794                             1
Name: Weight (Kilograms), Length: 4688, dtype: int64

Freight Cost (USD)
 Freight Included in Commodity Cost    1442
```

```
Invoiced Separately                    239
9736.1                                  36
6147.18                                 27
13398.06                                16
                                       ...
13740.74                                 1
1439                                     1
1726.5                                   1
4217.45                                  1
4485.25                                  1
Name: Freight Cost (USD), Length: 6733, dtype: int64
```

Detailed Description about the features:

PQ(Perfomance qualification):

It is the documented collection of activities necessary to demonstrate that an instrument consistently performs according to the specifications defined by the user and is appropriate for the intended use.

PO (Purchase Order) (or) SO (Shipping Order):

A purchase order (PO) is a document that represents an agreement with a vendor to buy goods or services

(or)

Shipping Order (SO) A document from the carrier that confirms that space for the cargo is booked onboard the vessel.

ASN(Advanced Shipment Notice) / DN(Delivery Note):

ASN: An advanced shipment notice (ASN) is an electronic data interchange (EDI) message sent from the shipper to the receiver prior to the departure of the shipment from the shipper's facility. The message includes complete information about the shipment and its contents.

DN: It provides a list of the products and quantity of the goods included in the deliver.

Country:

The country which the product is shipping from.

Managed by:

The particular shipment that is managed by the office i.e, Project Management Office.

We have four management offices,those are:

1)'PMO - US'

2)'South Africa Field Office'

3)'Haiti Field Office'

4)'Ethiopia Field Offic

Fullfill Via:

From RDC(Regional Distribution Center): A distribution center that distributes to provincial (state) users with strong radiation capacity and inventory preparation. This kind of distribution center has a large distribution scale.

Direct drop: Direct drop-shipping is where the retailer asks their suppliers or wholesalers to package and deliver orders directly to customers on their behalf.

Vendor INCO Term:

https://www.approvedforwarders.com/understanding-5-common-incoterms/#:~:text=Incoterms%20are%20a%20set%20of%20rules%20established%20by,Incoterms%

Incoterms specify who is responsible for paying for and managing the shipment, insurance, documentation, customs clearance, and other logistical activities. Incoterms:

EXW (Ex Works)

FCA (Free Carrier)

CPT (Carriage Paid To)

CIP (Carriage and Insurance Paid To)

DAP (Delivered at Place)

DDP (Delivered Duty Paid)

CIF(Cost,Insurance And Freight)

Shipment:

Means of transport.

PQ First Sent to Client Date:

On which date the Performance Qualification is sent to client.

We have information as Pre PQ process and dates.

PO Sent to Vendor Date:

On which date the Purchase Order sent to Vendor.

Scheduled Delivery Date:

The deivery will be delivered on that particular date.

Delivered to Client Date:

The date on which the product had delivered to client.

Delivery Recorded Date:

On which date the delivery happened is recorded.

Product Group:

Group of that particular product.

Sub Classification:

The classification of the product ccording to its use.

Vendor:

A company or a person offering for sale.(Trader)

Item Description:

Description about that item.

Molecule/ Test type:

Type of the molecule used in that Pharmaceutical product.

Dosage:

Quantity of the drug.

Dosage Form:

In which form that drug appears.Ex:- Tablet, Capsule, Injection etc.

Manufacturing Site:

A location where a Manufaturing operation is conducted.

First line Designation:

Attiribute to check the designation is first line or not.

Weight (Kilograms):

Weight of that particular packed product.

Freight Cost:

Freight costs are also known as freight charges or freight rates i.e the amount paid to a carrier company for the transportation of goods from the point of origin to an agreed location.

In [113…
```
df.head(1)
```

Out[113…

	ID	Project Code	PQ #	PO / SO #	ASN/DN #	Country	Managed By	Fulfill Via	Vendor INCO Term	Shipment Mode	PQ First Sent to Client Date	F
0	1	100-CI-T01	Pre-PQ Process	SCMS-4	ASN-8	Côte d'Ivoire	PMO - US	Direct Drop	EXW	Air	Pre-PQ Process	D C

Renaming the columns

In [114…
```
df.rename(columns={'PQ #':'Performance Qualification'},inplace=True)
```

In [115…
```
df.rename(columns={'PO / SO #':'Purchase(or)Shipping Order'},inplace=True)
```

In [116…
```
df.rename(columns={'ASN/DN #':'Advanced Shipment Notice'},inplace=True)
```

In [117…
```
df.head(1)
```

Out[117…

	ID	Project Code	Performance Qualification	Purchase(or)Shipping Order	Advanced Shipment Notice	Country	Managed By	Fulfill Via	Vendor INCO Term
0	1	100-CI-T01	Pre-PQ Process	SCMS-4	ASN-8	Côte d'Ivoire	PMO - US	Direct Drop	EXW

Numerical Features

In [118...

```
num_feat=df.select_dtypes(include='number')
print('The Numerical Features of the data are:','\n',num_feat.columns)
```

```
The Numerical Features of the data are:
 Index(['ID', 'Unit of Measure (Per Pack)', 'Line Item Quantity',
       'Line Item Value', 'Pack Price', 'Unit Price',
       'Line Item Insurance (USD)'],
      dtype='object')
```

Handling Missing Values

In [119...

```
df.isnull().sum()
```

Out[119...

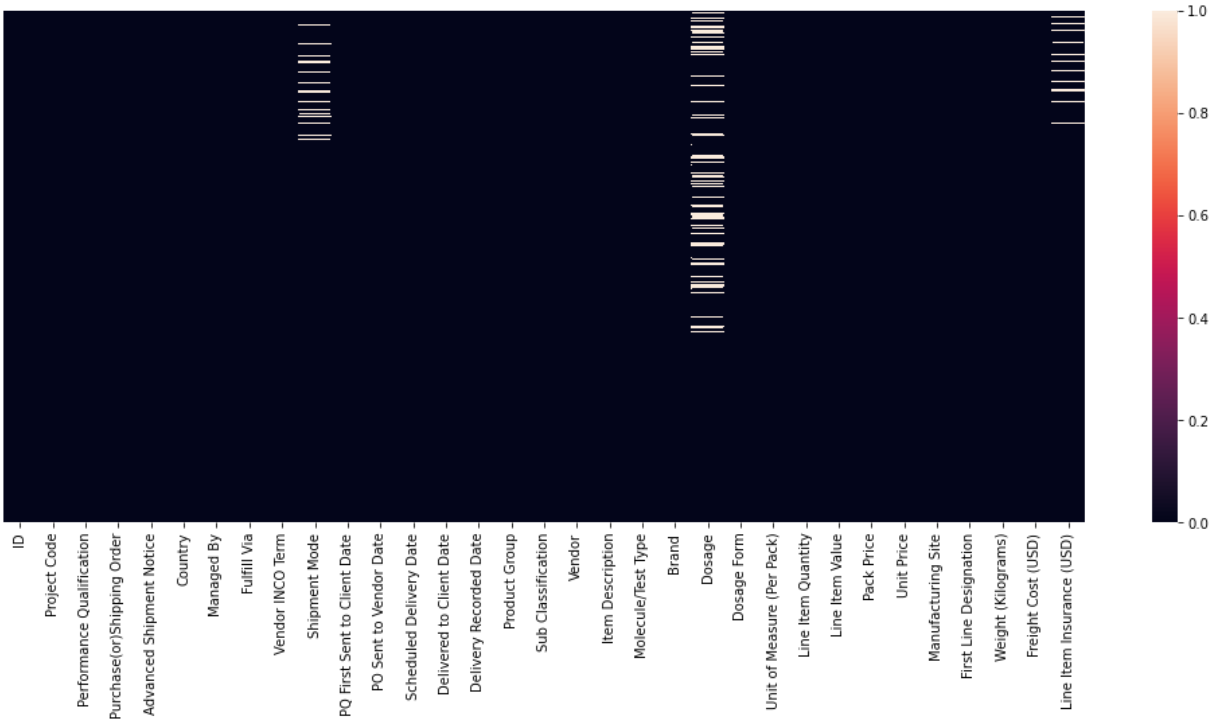
```
ID                                0
Project Code                      0
Performance Qualification         0
Purchase(or)Shipping Order        0
Advanced Shipment Notice          0
Country                           0
Managed By                        0
Fulfill Via                       0
Vendor INCO Term                  0
Shipment Mode                   360
PQ First Sent to Client Date      0
PO Sent to Vendor Date            0
Scheduled Delivery Date           0
Delivered to Client Date          0
Delivery Recorded Date            0
Product Group                     0
Sub Classification                0
Vendor                            0
Item Description                  0
Molecule/Test Type                0
Brand                             0
Dosage                         1736
Dosage Form                       0
Unit of Measure (Per Pack)        0
Line Item Quantity                0
Line Item Value                   0
Pack Price                        0
Unit Price                        0
Manufacturing Site                0
First Line Designation            0
Weight (Kilograms)                0
Freight Cost (USD)                0
```

```
Line Item Insurance (USD)        287
dtype: int64
```

Visualizing the Null values

In [120…

```
sns.heatmap(df.isnull(), yticklabels=False);
```



Observation:

Here we can see that null values are present in the columns Shipment Mode, Dosage and Line Insurance(USD).

Handling Null values in "Doasge" category.

In [121…

```
df1=df[df['Dosage'].isnull()]
df1
```

Out[121…

	ID	Project Code	Performance Qualification	Purchase(or)Shipping Order	Advanced Shipment Notice	Country	Managed By	Fulfill Via	V
0	1	100-CI-T01	Pre-PQ Process	SCMS-4	ASN-8	Côte d'Ivoire	PMO - US	Direct Drop	
2	4	100-CI-T01	Pre-PQ Process	SCMS-20	ASN-14	Côte d'Ivoire	PMO - US	Direct Drop	

	ID	Project Code	Performance Qualification	Purchase(or)Shipping Order	Advanced Shipment Notice	Country	Managed By	Fulfill Via	V
11	61	110-ZM-T01	Pre-PQ Process	SCMS-226	ASN-137	Zambia	PMO - US	Direct Drop	
12	62	102-NG-T01	Pre-PQ Process	SCMS-230	ASN-144	Nigeria	PMO - US	Direct Drop	
14	65	106-HT-T01	Pre-PQ Process	SCMS-274	ASN-162	Haiti	PMO - US	Direct Drop	
...	...	...	...	...	...	...	...	...	
8901	85082	102-SD-T30	FPQ-3812	SO-34550	DN-1624	Sudan	PMO - US	From RDC	
9564	85865	110-ZM-T30	FPQ-9515	SO-41950	DN-2681	Zambia	PMO - US	From RDC	
9734	86064	101-KE-T30	FPQ-13822	SO-48080	DN-3591	Kenya	PMO - US	From RDC	
9863	86258	102-SD-T30	FPQ-3812	SO-34550	DN-1624	Sudan	PMO - US	From RDC	
10055	86457	109-TZ-T30	FPQ-9275	SO-41460	DN-2583	Tanzania	PMO - US	From RDC	

1736 rows × 33 columns

Attribute 'Dosage' has 1736 missing values

In [122...
```
df1.shape
```

Out[122... (1736, 33)

In [123...
```
df['Product Group'].value_counts()
```

Out[123…
```
ARV     8550
HRDT    1728
ANTM      22
ACT       16
MRDT       8
Name: Product Group, dtype: int64
```

In [124…
```
new_df=df[df['Product Group']=='HRDT']
new_df
```

Out[124…

	ID	Project Code	Performance Qualification	Purchase(or)Shipping Order	Advanced Shipment Notice	Country	Managed By	Fulfill Via	V
0	1	100-CI-T01	Pre-PQ Process	SCMS-4	ASN-8	Côte d'Ivoire	PMO - US	Direct Drop	
2	4	100-CI-T01	Pre-PQ Process	SCMS-20	ASN-14	Côte d'Ivoire	PMO - US	Direct Drop	
11	61	110-ZM-T01	Pre-PQ Process	SCMS-226	ASN-137	Zambia	PMO - US	Direct Drop	
12	62	102-NG-T01	Pre-PQ Process	SCMS-230	ASN-144	Nigeria	PMO - US	Direct Drop	
14	65	106-HT-T01	Pre-PQ Process	SCMS-274	ASN-162	Haiti	PMO - US	Direct Drop	
...	...	...	...	...	...	...	...	...	
8901	85082	102-SD-T30	FPQ-3812	SO-34550	DN-1624	Sudan	PMO - US	From RDC	
9564	85865	110-ZM-T30	FPQ-9515	SO-41950	DN-2681	Zambia	PMO - US	From RDC	
9734	86064	101-KE-T30	FPQ-13822	SO-48080	DN-3591	Kenya	PMO - US	From RDC	

	ID	Project Code	Performance Qualification	Purchase(or)Shipping Order	Advanced Shipment Notice	Country	Managed By	Fulfill Via	V
9863	86258	102-SD-T30	FPQ-3812	SO-34550	DN-1624	Sudan	PMO - US	From RDC	
10055	86457	109-TZ-T30	FPQ-9275	SO-41460	DN-2583	Tanzania	PMO - US	From RDC	

1728 rows × 33 columns

In [125…
```
new_df['Dosage Form'].value_counts()
```

Out[125…
```
Test kit               1567
Test kit - Ancillary    161
Name: Dosage Form, dtype: int64
```

Observation:

Here we can observe that the null values in 'Dosage' column are given for entire 'Test kit' and 'Test kit-Ancellary' categories in 'Dosage form' column which are all belonging to HRDT and MRDT 'Product groups' and we cannot describe the Dosage weight for Test kits because the 'Test Kit' actually means that collection of medicines for First Aid. We can describe the Dosage weight for it as 'No weight Measurement'.

In [126…
```
df['Dosage']=df['Dosage'].fillna(df['Dosage'].isnull().values=='No Weight Measuremen
```

Handling Null values in 'shipment Mode'

Data had 360 null values in the column 'Shipment Mode' that can filled with Highly repeated value(i.e Mode)

In [127…
```
df['Shipment Mode'].value_counts()
```

Out[127…
```
Air            6113
Truck          2830
Air Charter     650
Ocean           371
Name: Shipment Mode, dtype: int64
```

Observation:

Here we can see that the 'Air' shipment mode is high. Therefore we can replace the null values with this this category.

In [128…
```
df['Shipment Mode']=df['Shipment Mode'].replace(np.nan,str(df['Shipment Mode'].mode(
```

Handling Null values in "Line Insurance(USD)"

There are 287 null values in the'Line Item Insurance (USD)' column which is a numrical variable.

Observation:

The column 'Line Item Insurance (USD)' is a numerical variable,hence the null values can be replaced with the mean of all values in that column

In [129…
```
df['Line Item Insurance (USD)']=df['Line Item Insurance (USD)'].fillna(df['Line Item
```

Missing values - Cross check

In [130…
```
df.isnull().sum()
```

Out[130…
```
ID                                0
Project Code                      0
Performance Qualification         0
Purchase(or)Shipping Order        0
Advanced Shipment Notice          0
Country                           0
Managed By                        0
Fulfill Via                       0
Vendor INCO Term                  0
Shipment Mode                     0
PQ First Sent to Client Date      0
PO Sent to Vendor Date            0
Scheduled Delivery Date           0
Delivered to Client Date          0
Delivery Recorded Date            0
Product Group                     0
Sub Classification                0
Vendor                            0
Item Description                  0
Molecule/Test Type                0
Brand                             0
Dosage                            0
Dosage Form                       0
Unit of Measure (Per Pack)        0
Line Item Quantity                0
Line Item Value                   0
Pack Price                        0
Unit Price                        0
Manufacturing Site                0
First Line Designation            0
Weight (Kilograms)                0
Freight Cost (USD)                0
Line Item Insurance (USD)         0
dtype: int64
```

In [131…
```
#Visualising again.
sns.heatmap(df.isnull(),yticklabels=False);
```

Duplicated records

In [132…
```
df.duplicated().sum()
```

Out[132… 0

Observation:

There are no duplicate records in the given data

Outliers

In [133…
```
df1=df.describe(percentiles=[0.01,0.05,0.1,0.25,0.5,0.75,0.90,0.95,0.99])
df1
```

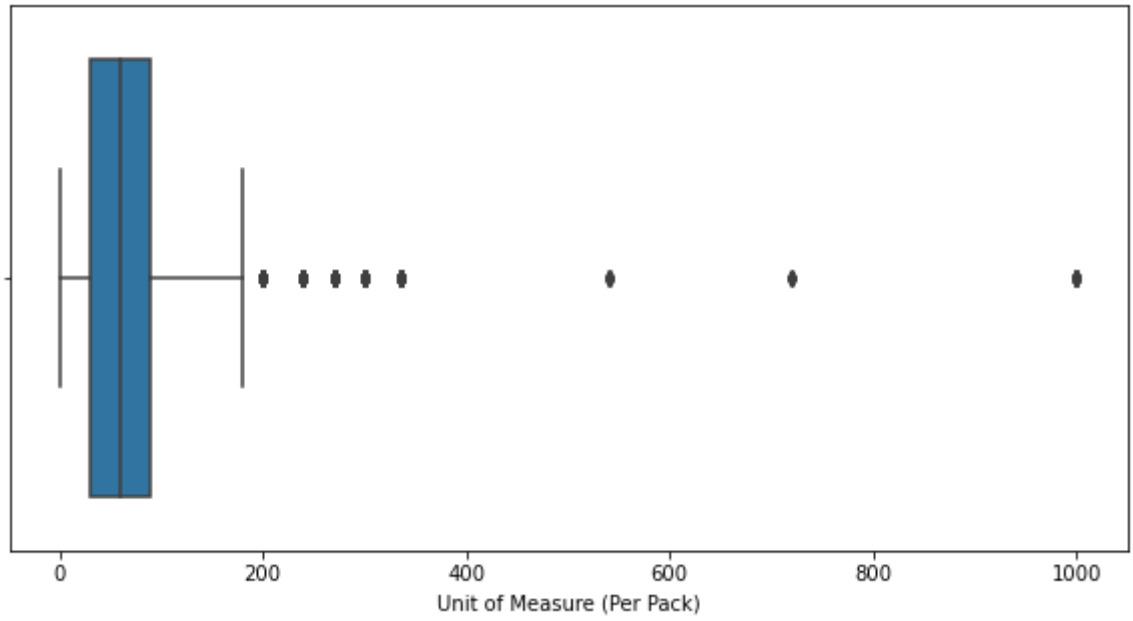
Out[133…

	ID	Unit of Measure (Per Pack)	Line Item Quantity	Line Item Value	Pack Price	Unit Price	Lir Ins
count	10324.000000	10324.000000	10324.000000	1.032400e+04	10324.000000	10324.000000	10324.
mean	51098.968229	77.990895	18332.534870	1.576506e+05	21.910241	0.611701	240.
std	31944.332496	76.579764	40035.302961	3.452921e+05	45.609223	3.275808	493.
min	1.000000	1.000000	1.000000	0.000000e+00	0.000000	0.000000	0.
1%	1078.230000	1.000000	3.000000	1.587680e+01	0.390000	0.010000	0.
5%	5151.200000	20.000000	17.000000	1.925755e+02	1.900000	0.010000	0.
10%	10547.300000	30.000000	50.000000	6.081200e+02	2.240000	0.040000	0.
25%	12795.750000	30.000000	408.000000	4.314593e+03	4.120000	0.080000	7.
50%	57540.500000	60.000000	3000.000000	3.047147e+04	9.300000	0.160000	52.
75%	83648.250000	90.000000	17039.750000	1.664471e+05	23.592500	0.470000	241.
90%	85541.700000	180.000000	53267.000000	4.374878e+05	69.940000	0.890000	669.

	ID	Unit of Measure (Per Pack)	Line Item Quantity	Line Item Value	Pack Price	Unit Price	Lir Ins
95%	86167.850000	240.000000	90951.550000	7.028310e+05	80.000000	1.600000	1061.
99%	86651.770000	300.000000	186888.780000	1.592935e+06	139.000000	5.000000	2419.
max	86823.000000	1000.000000	619999.000000	5.951990e+06	1345.640000	238.650000	7708.

In [134…

```
plt.figure(figsize=(10,5))
sns.boxplot('Unit of Measure (Per Pack)',data=df);
```



In [135…

```
A=df['Unit of Measure (Per Pack)'].quantile(0.25)
B=df['Unit of Measure (Per Pack)'].quantile(0.75)

#Inter Quartile Range
IQR=B-A
print(IQR)
lower_limit=A-1.5*IQR
upper_limit=B+1.5*IQR
print(lower_limit,upper_limit)
```

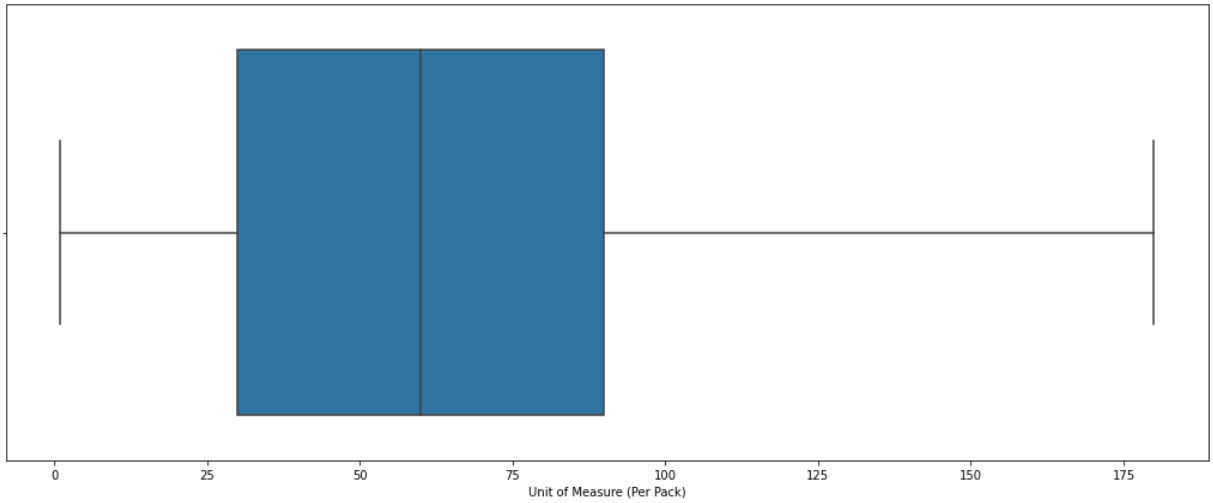
```
60.0
-60.0 180.0
```

In [136…

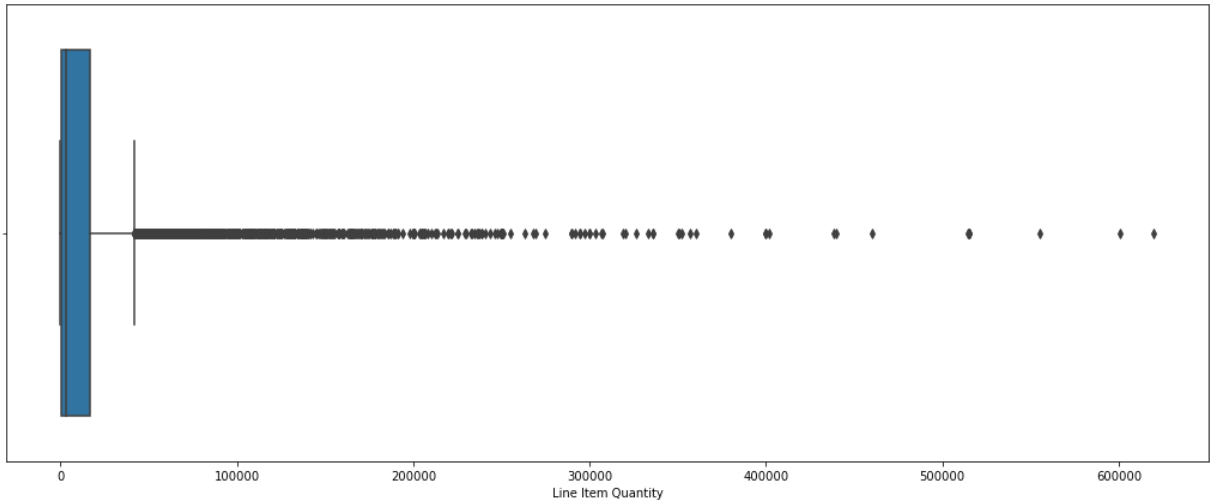
```
df['Unit of Measure (Per Pack)']=np.where(df['Unit of Measure (Per Pack)']>180,180,d
```

In [137…

```
sns.boxplot('Unit of Measure (Per Pack)',data=df);
```



In [138…
```
sns.boxplot('Line Item Quantity',data=df);
```

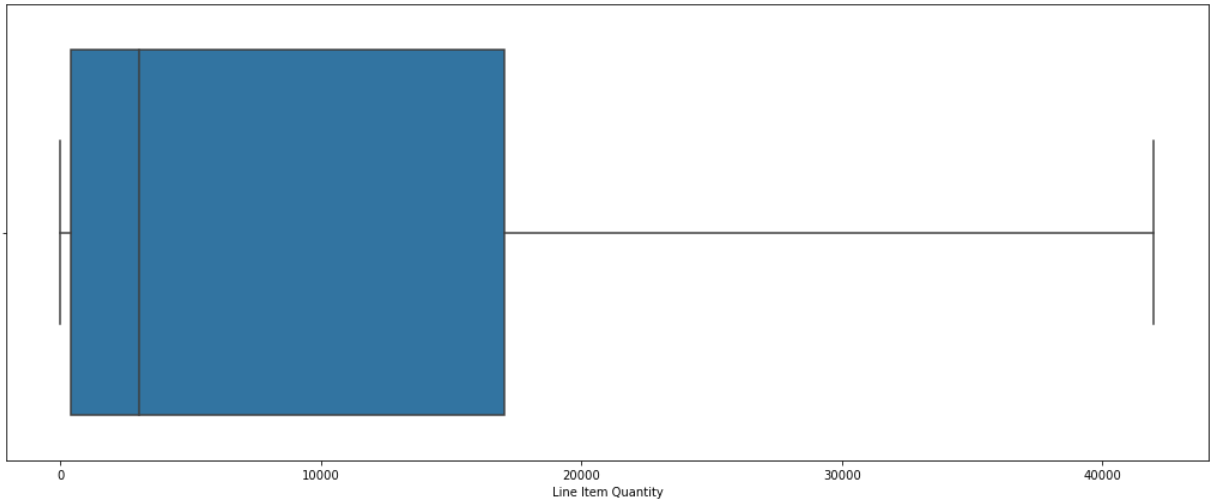


In [139…
```
A=df['Line Item Quantity'].quantile(0.25)
B=df['Line Item Quantity'].quantile(0.75)
IQR=B-A
print(IQR)
lower_limit=A-1.5*IQR
upper_limit=B+1.5*IQR
print(lower_limit,upper_limit)
```

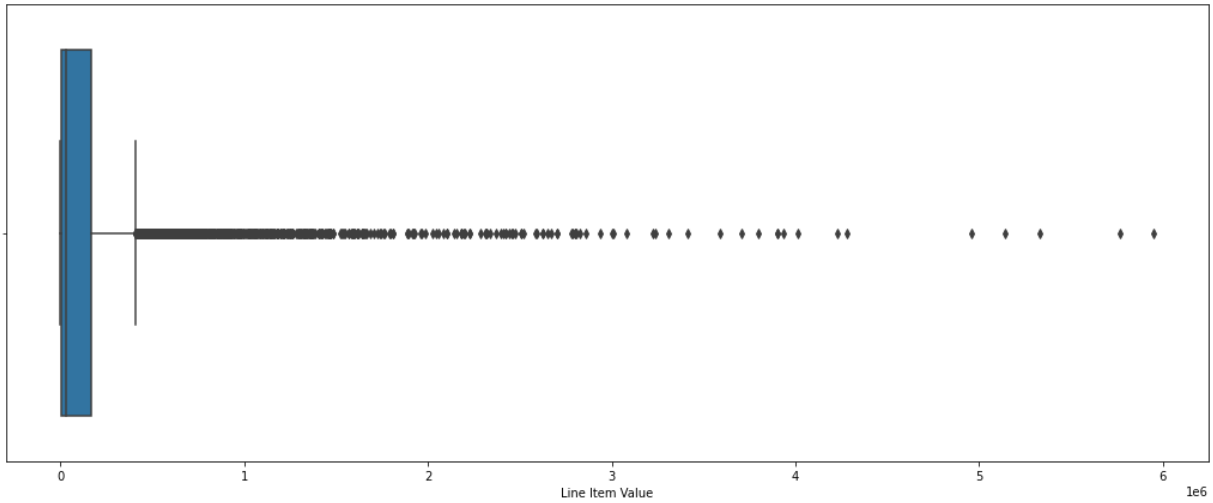
```
16631.75
-24539.625 41987.375
```

In [140…
```
df['Line Item Quantity']=np.where(df['Line Item Quantity']>41987,41987,df['Line Item
```

In [141…
```
sns.boxplot('Line Item Quantity',data=df);
```



In [142…
```
sns.boxplot('Line Item Value',data=df);
```

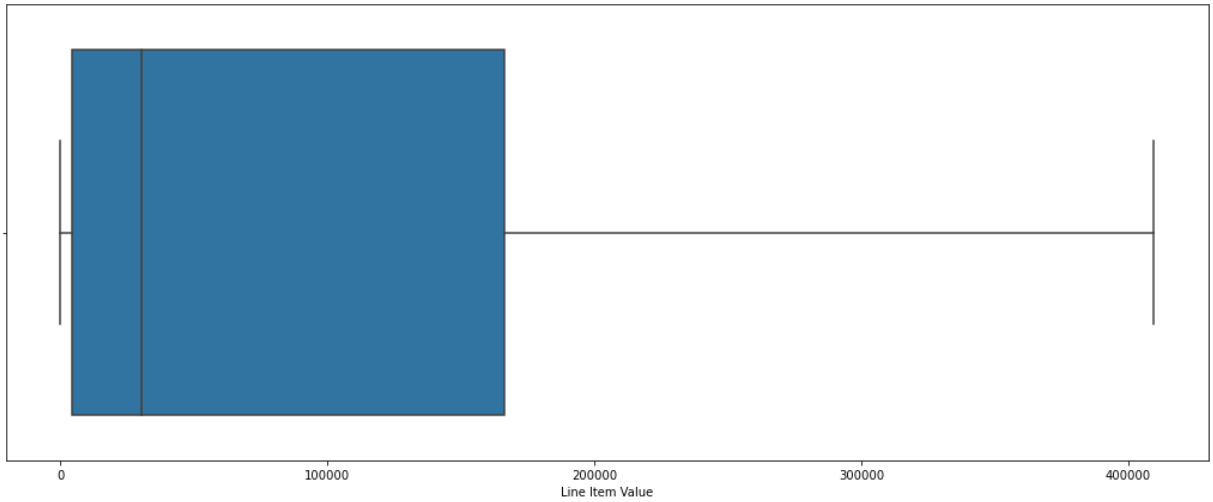


In [143…
```
#Extreme Outliers for 'Line Item Value'
A=df['Line Item Value'].quantile(0.25)
B=df['Line Item Value'].quantile(0.75)
IQR=B-A
print(IQR)
lower_limit=A-1.5*IQR
upper_limit=B+1.5*IQR
print(lower_limit,upper_limit)
```

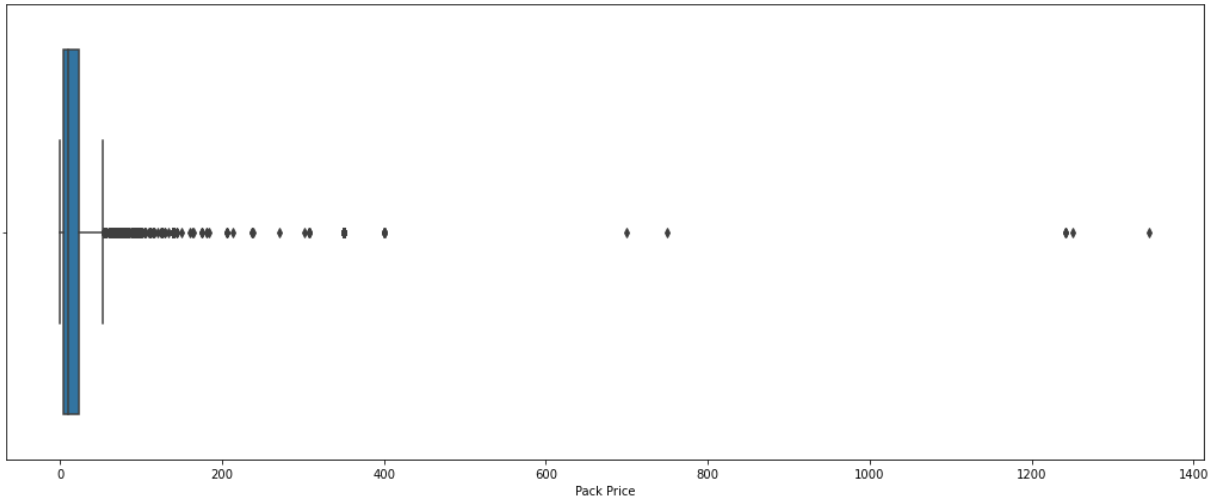
```
162132.5475
-238884.22874999998 409645.96124999993
```

In [144…
```
df['Line Item Value']=np.where(df['Line Item Value']>409645,409645,df['Line Item Val
```

In [145…
```
sns.boxplot('Line Item Value',data=df);
```



In [146…
```
sns.boxplot('Pack Price',data=df);
```

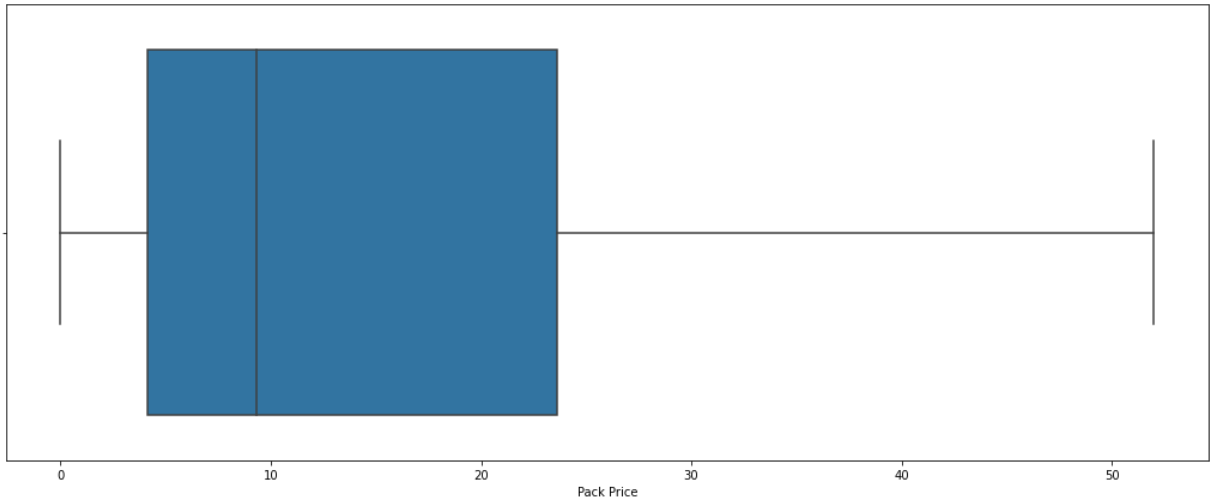


In [147…
```
A=df['Pack Price'].quantile(0.25)
B=df['Pack Price'].quantile(0.75)
IQR=B-A
print(IQR)
lower_limit=A-1.5*IQR
upper_limit=B+1.5*IQR
print(lower_limit,upper_limit)
```

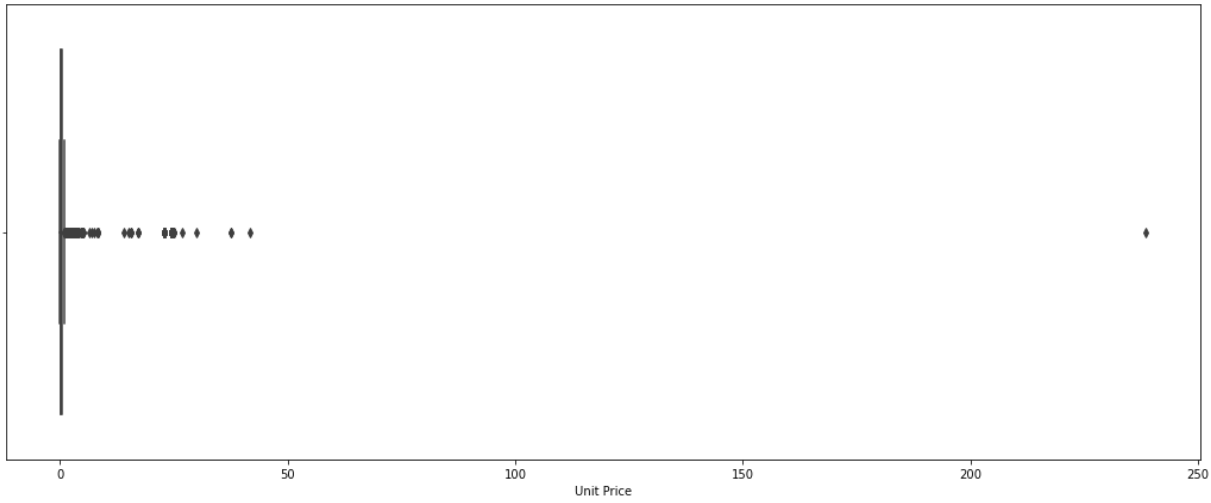
```
19.4725
-25.08875 52.80125
```

In [148…
```
df['Pack Price']=np.where(df['Pack Price']>52,52,df['Pack Price'])
```

In [149…
```
sns.boxplot('Pack Price',data=df);
```



In [150…
```
sns.boxplot('Unit Price',data=df);
```

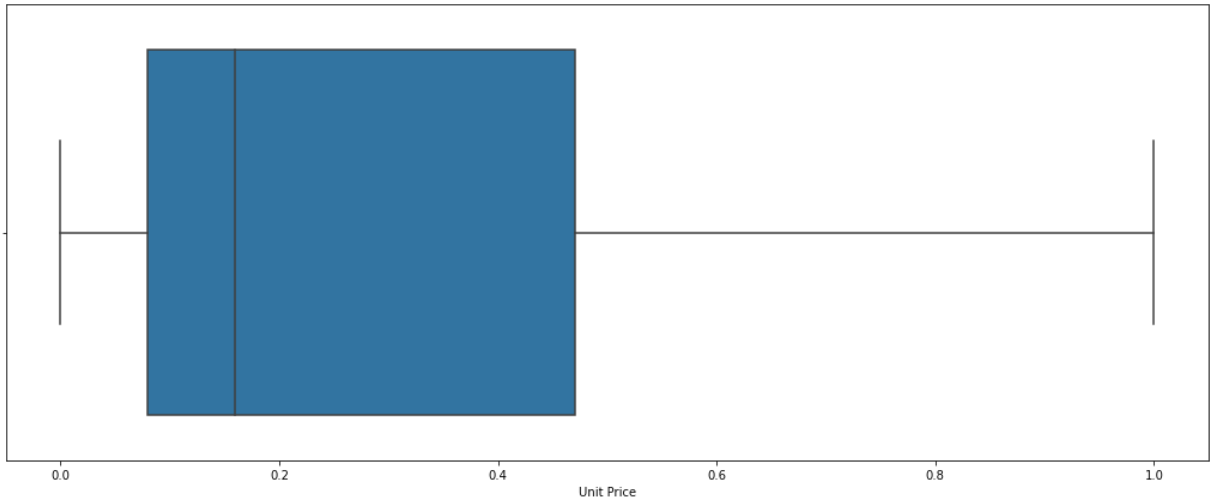


In [151…
```
A=df['Unit Price'].quantile(0.25)
B=df['Unit Price'].quantile(0.75)
IQR=B-A
print(IQR)
lower_limit=A-1.5*IQR
upper_limit=B+1.5*IQR
print(lower_limit,upper_limit)
```

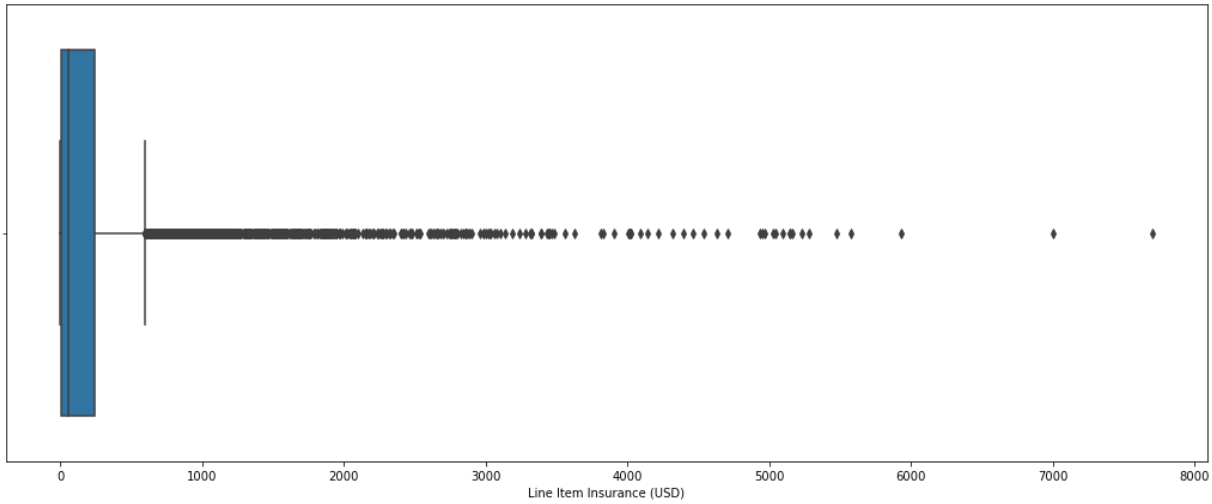
```
0.38999999999999996
-0.505 1.055
```

In [152…
```
df['Unit Price']=np.where(df['Unit Price']>1,1,df['Unit Price'])
```

In [153…
```
sns.boxplot('Unit Price',data=df);
```



In [154…
```
sns.boxplot('Line Item Insurance (USD)',data=df);
```

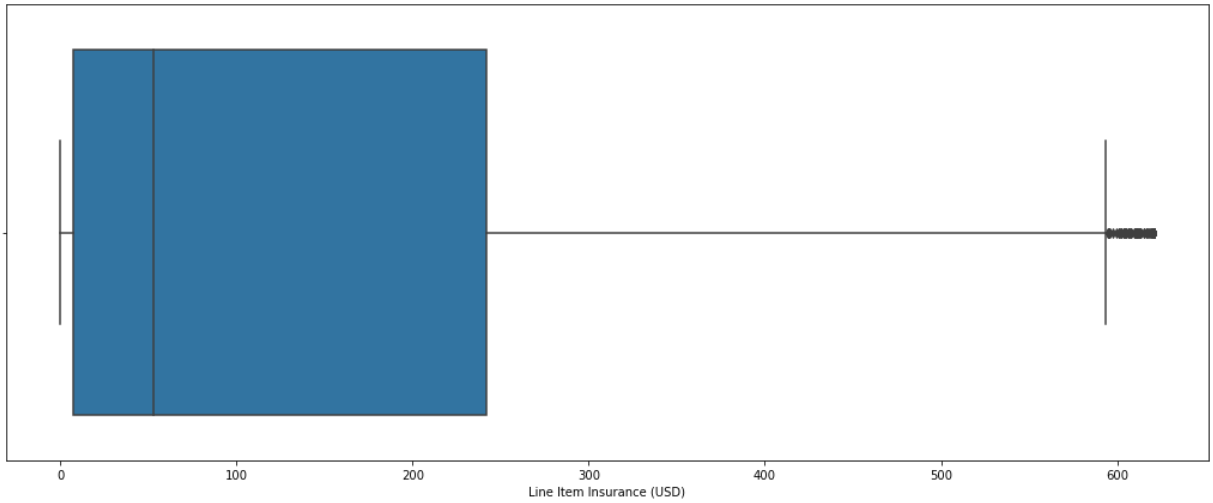


In [155…
```
A=df['Line Item Insurance (USD)'].quantile(0.25)
B=df['Line Item Insurance (USD)'].quantile(0.75)
IQR=B-A
print(IQR)
lower_limit=A-1.5*IQR
upper_limit=B+1.5*IQR
print(lower_limit,upper_limit)
```

```
234.72
-345.05 593.8299999999999
```

In [156…
```
df['Line Item Insurance (USD)']=np.where(df['Line Item Insurance (USD)']>621,621,df[
```

In [157…
```
sns.boxplot('Line Item Insurance (USD)',data=df);
```



In [158...
```
df.describe()
```

Out[158...

	ID	Unit of Measure (Per Pack)	Line Item Quantity	Line Item Value	Pack Price	Unit Price	Lin Ins
count	10324.000000	10324.000000	10324.000000	10324.000000	10324.000000	10324.000000	10324.(
mean	51098.968229	69.297172	11163.128729	105691.099440	16.215625	0.315993	163.7
std	31944.332496	46.417603	14990.979423	139246.742744	16.190944	0.323232	209.(
min	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.(
25%	12795.750000	30.000000	408.000000	4314.592500	4.120000	0.080000	7.(
50%	57540.500000	60.000000	3000.000000	30471.465000	9.300000	0.160000	52.9
75%	83648.250000	90.000000	17039.750000	166447.140000	23.592500	0.470000	241.7
max	86823.000000	180.000000	41987.000000	409645.000000	52.000000	1.000000	621.(

In [159...
```
df.head()
```

Out[159...

	ID	Project Code	Performance Qualification	Purchase(or)Shipping Order	Advanced Shipment Notice	Country	Managed By	Fulfill Via	Vendor INCO Term
0	1	100-CI-T01	Pre-PQ Process	SCMS-4	ASN-8	Côte d'Ivoire	PMO - US	Direct Drop	EXW
1	3	108-VN-T01	Pre-PQ Process	SCMS-13	ASN-85	Vietnam	PMO - US	Direct Drop	EXW

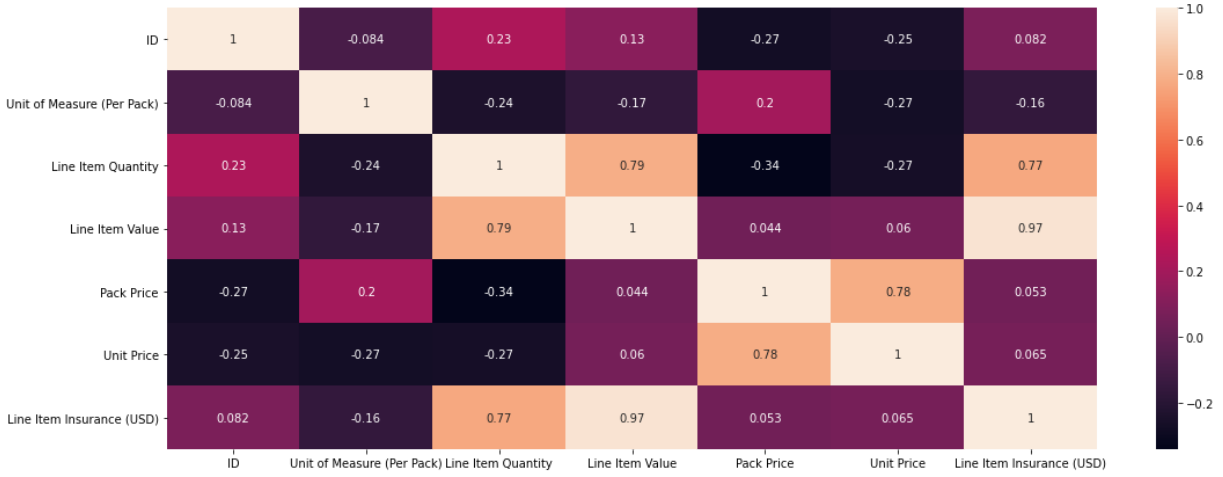
	ID	Project Code	Performance Qualification	Purchase(or)Shipping Order	Advanced Shipment Notice	Country	Managed By	Fulfill Via	Vendor INCO Term
2	4	100-CI-T01	Pre-PQ Process	SCMS-20	ASN-14	Côte d'Ivoire	PMO - US	Direct Drop	FCA
3	15	108-VN-T01	Pre-PQ Process	SCMS-78	ASN-50	Vietnam	PMO - US	Direct Drop	EXW
4	16	108-VN-T01	Pre-PQ Process	SCMS-81	ASN-55	Vietnam	PMO - US	Direct Drop	EXW

In [ ]:

Correlation Matrix

In [160…

```
corr=df.corr()
sns.heatmap(corr,annot=True);
```



Observation:

Here we can observe that the columns "Line Item Insurance(USD)", "Line Item Value" and "Line Item Quantity" are highly correlated with each other,so any one of them is considered.

Line Item:

Line Item is a code taht is given to each item with respect to the coveyor belt or the machine it is created on.

Line Item Quantity:

Line item Quantity is supposed to be the number of items created in that machine.

In [161...
```
df.drop(['Line Item Insurance (USD)' ,'Line Item Quantity'],axis=1,inplace=True)
```

In [162...
```
#VIsualise the data and check whther the columns are present or not.
df.head(2)
```

Out[162...

	ID	Project Code	Performance Qualification	Purchase(or)Shipping Order	Advanced Shipment Notice	Country	Managed By	Fulfill Via	Vendor INCO Term
0	1	100-CI-T01	Pre-PQ Process	SCMS-4	ASN-8	Côte d'Ivoire	PMO - US	Direct Drop	EXW
1	3	108-VN-T01	Pre-PQ Process	SCMS-13	ASN-85	Vietnam	PMO - US	Direct Drop	EXW

Observation:

Here we can also see that the columns 'Delivered to Client Date' and 'Delivered Recorded Date' represents the dates which are same,so any one of the both columns is excluded .

In [163...
```
df.drop('Delivery Recorded Date',axis=1,inplace=True)
```

In [164...
```
df.head()
```

Out[164...

	ID	Project Code	Performance Qualification	Purchase(or)Shipping Order	Advanced Shipment Notice	Country	Managed By	Fulfill Via	Vendor INCO Term
0	1	100-CI-T01	Pre-PQ Process	SCMS-4	ASN-8	Côte d'Ivoire	PMO - US	Direct Drop	EXW
1	3	108-VN-T01	Pre-PQ Process	SCMS-13	ASN-85	Vietnam	PMO - US	Direct Drop	EXW
2	4	100-CI-T01	Pre-PQ Process	SCMS-20	ASN-14	Côte d'Ivoire	PMO - US	Direct Drop	FCA

	ID	Project Code	Performance Qualification	Purchase(or)Shipping Order	Advanced Shipment Notice	Country	Managed By	Fulfill Via	Vendor INCO Term
3	15	108-VN-T01	Pre-PQ Process	SCMS-78	ASN-50	Vietnam	PMO - US	Direct Drop	EXW
4	16	108-VN-T01	Pre-PQ Process	SCMS-81	ASN-55	Vietnam	PMO - US	Direct Drop	EXW

Conclusion:

Here our task which we had performed is Data Cleaning or Data Wrangling that consists the process of removing duplicate or irrelevant observations, removing unwanted observations from Dataset, excluding the outliers, handling Missing values and validation. And further process has to be done for performing the model creation and deployment is Data exploration and Visualization , Feature Engineering and Feature Scaling and later the model will be developed.

In [ ]: