In [33]:

```
import numpy as np
import pandas as pd
```

In [34]:

```
# Reading a .csv file
```

In [35]:

```
df = pd.read_csv('helpdesk_tickets.csv')
```

In [36]:

```
# take a data information what types, columns and rows
```

In [37]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 565 entries, 0 to 564
Data columns (total 27 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   Ticket Number                565 non-null    int64  
 1   Date Created                 565 non-null    object 
 2   Subject                      565 non-null    object 
 3   From                         565 non-null    object 
 4   From Email                   565 non-null    object 
 5   Priority                     565 non-null    object 
 6   Department                   565 non-null    object 
 7   Type                         563 non-null    object 
 8   Source                       565 non-null    object 
 9   Current Status               565 non-null    object 
 10  Last Updated                 565 non-null    object 
 11  Due Date                     460 non-null    object 
 12  Overdue                      565 non-null    int64  
 13  Answered                     565 non-null    int64  
 14  Agent Assigned               565 non-null    object 
 15  Team Lead                    0 non-null      float64
 16  Team Assigned                565 non-null    object 
 17  Thread Count                 565 non-null    int64  
 18  Attachment Count             565 non-null    int64  
 19  Category                     289 non-null    object 
 20  Issue Origin                 301 non-null    object 
 21  Select Ticket Status Update  561 non-null    object 
 22  Unnamed: 22                  0 non-null      float64
 23  SR No.                       0 non-null      float64
 24  Corrective Actions           0 non-null      float64
 25  Preventive Actions           0 non-null      float64
 26  Closed By                    0 non-null      float64
dtypes: float64(6), int64(5), object(16)
memory usage: 119.3+ KB
```

In [ ]:

```
# Checking the names of the columns
```

In [38]:

```
df.columns
```

Out[38]:

```
Index(['Ticket Number', 'Date Created', 'Subject', 'From', 'From Email',
       'Priority', 'Department', 'Type', 'Source', 'Current Status',
       'Last Updated', 'Due Date', 'Overdue', 'Answered', 'Agent Assigned',
       'Team Lead', 'Team Assigned', 'Thread Count', 'Attachment Count',
       'Category', 'Issue Origin', 'Select Ticket Status Update',
       'Unnamed: 22', 'SR No.', 'Corrective Actions', 'Preventive Actions',
       'Closed By'],
      dtype='object')
```

In [ ]:

```
# Quic look over first 3 data rows
```

In [39]:

```
df.head(3)
```

Out[39]:

	Ticket Number	Date Created	Subject	From	From Email	Priority	Department
0	111636	3/12/21 9:57	Error Displaying in Different Module	Jasper John	jasper.john@gmail.com	Emergency	SAP JDE Support Department
1	111632	3/10/21 16:21	Approval Workflow Error	Erick White	ewhite@yahoo.com	Emergency	SAP JDE Support Department
2	111621	2/22/21 12:08	Public IP Trusted Certificate Authority Error	Tomi Yamamoto	tyamamoto@gmail.com	Emergency	Internal Technical Department

3 rows × 27 columns

In [ ]:

```
# Make data discription
```

In [40]:

```
df.describe()
```

Out[40]:

	Ticket Number	Overdue	Answered	Team Lead	Thread Count	Attachment Count	Unnamed: 22
count	565.000000	565.000000	565.000000	0.0	565.000000	565.000000	0.0
mean	114969.304425	0.074336	0.929204	NaN	13.534513	2.695575	NaN
std	44957.649446	0.262550	0.256712	NaN	15.392073	4.039178	NaN
min	2.000000	0.000000	0.000000	NaN	1.000000	0.000000	NaN
25%	111240.000000	0.000000	1.000000	NaN	5.000000	0.000000	NaN
50%	111377.000000	0.000000	1.000000	NaN	9.000000	2.000000	NaN
75%	111523.000000	0.000000	1.000000	NaN	16.000000	3.000000	NaN
max	694809.000000	1.000000	1.000000	NaN	179.000000	43.000000	NaN

In [ ]:

```
# Ticket Number have irrelevant data min = 2, mean = 114969  and max = 694
# The data has 6 empty columns
# Chek for duplicates
```

In [41]:

```
df.duplicated()
```

Out[41]:

```
0      False
1      False
2      False
3      False
4      False
       ...
560     True
561     True
562     True
563     True
564     True
Length: 565, dtype: bool
```

In [ ]:

```
# data has duplicates
```

In [42]:

```
df.duplicated().sum()
```

Out[42]:

```
16
```

In [43]:

```
dfremoveduplicates = df.drop_duplicates()
```

In [ ]:

```
# check again the new data after removed duplicates
```

In [44]:

```
dfremoveduplicates.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 549 entries, 0 to 548
Data columns (total 27 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   Ticket Number                549 non-null    int64  
 1   Date Created                 549 non-null    object 
 2   Subject                      549 non-null    object 
 3   From                         549 non-null    object 
 4   From Email                   549 non-null    object 
 5   Priority                     549 non-null    object 
 6   Department                   549 non-null    object 
 7   Type                         547 non-null    object 
 8   Source                       549 non-null    object 
 9   Current Status               549 non-null    object 
 10  Last Updated                 549 non-null    object 
 11  Due Date                     445 non-null    object 
 12  Overdue                      549 non-null    int64  
 13  Answered                     549 non-null    int64  
 14  Agent Assigned               549 non-null    object 
 15  Team Lead                    0 non-null      float64
 16  Team Assigned                549 non-null    object 
 17  Thread Count                 549 non-null    int64  
 18  Attachment Count             549 non-null    int64  
 19  Category                     281 non-null    object 
 20  Issue Origin                 293 non-null    object 
 21  Select Ticket Status Update  545 non-null    object 
 22  Unnamed: 22                  0 non-null      float64
 23  SR No.                       0 non-null      float64
 24  Corrective Actions           0 non-null      float64
 25  Preventive Actions           0 non-null      float64
 26  Closed By                    0 non-null      float64
dtypes: float64(6), int64(5), object(16)
memory usage: 120.1+ KB
```

In [ ]:

```
# Remove the outlier of Ticket Number columns
```

In [51]:

```
dfremoveduplicates = dfremoveduplicates[(dfremoveduplicates['Ticket Number'
```

In [ ]:

```
# Remove the 6 empty columns from data
```

In [52]:

```
dfcleanmissdata = dfremoveduplicates.drop(['Team Lead','Unnamed: 22', 'SR N
       'Closed By'], axis=1)
```

In [53]:

```
dfcleanmissdata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 543 entries, 0 to 546
Data columns (total 21 columns):
 #   Column                       Non-Null Count  Dtype 
---  ------                       --------------  ----- 
 0   Ticket Number                543 non-null    int64 
 1   Date Created                 543 non-null    object
 2   Subject                      543 non-null    object
 3   From                         543 non-null    object
 4   From Email                   543 non-null    object
 5   Priority                     543 non-null    object
 6   Department                   543 non-null    object
 7   Type                         543 non-null    object
 8   Source                       543 non-null    object
 9   Current Status               543 non-null    object
 10  Last Updated                 543 non-null    object
 11  Due Date                     439 non-null    object
 12  Overdue                      543 non-null    int64 
 13  Answered                     543 non-null    int64 
 14  Agent Assigned               543 non-null    object
 15  Team Assigned                543 non-null    object
 16  Thread Count                 543 non-null    int64 
 17  Attachment Count             543 non-null    int64 
 18  Category                     280 non-null    object
 19  Issue Origin                 292 non-null    object
 20  Select Ticket Status Update  543 non-null    object
dtypes: int64(5), object(16)
memory usage: 93.3+ KB
```

In [49]:

```
dfcleanmissdata
```

Out[49]:

	Ticket Number	Date Created	Subject	From	From Email	Priority	Depa
0	111636	3/12/21 9:57	Error Displaying in Different Module	Jasper John	jasper.john@gmail.com	Emergency	S Dep
1	111632	3/10/21 16:21	Approval Workflow Error	Erick White	ewhite@yahoo.com	Emergency	S Dep
2	111621	2/22/21 12:08	Public IP Trusted Certificate Authority Error	Tomi Yamamoto	tyamamoto@gmail.com	Emergency	T Dep
3	111608	2/15/21 11:41	JDE Slowdown	Riza Richardson	rrichardson@mailinator.com	Emergency	S Dep
4	111596	1/22/21 10:58	JDE Slow Down	Riza Richardson	rrichardson@mailinator.com	Emergency	S Dep
...	...	...	...	...	...	...	
542	111236	6/19/19 13:00	SAP Dev Instance not Accessible	Aurora Miller	aurora.miller@outlook.com	Low	S Dep
543	111202	5/15/19 13:07	Can't access Dev. Instance	Jasper John	jasper.john@gmail.com	Low	S Dep
544	111205	5/17/19 15:33	Faculty Roles for Viewing Query Report	Jasper John	jasper.john@gmail.com	Low	S Dep
545	111128	2/15/19 16:42	Component Interface based Web Services ERROR	John Brown	jbrown@outlook.com	Low	S Dep
546	111124	2/8/19 9:46	WGET.SH Error Encountered	John Brown	jbrown@outlook.com	Low	S Dep

543 rows × 21 columns

In [ ]:

```
# save the new cleaning data
```

In [50]:

```
dfcleanmissdata.to_csv('new_helpdesk_tickets.csv',index=False)
```

In [ ]: