Table extraction using Tabula

In [25]:
```
import tabula
import pandas as pd
import numpy as np
from IPython.display import display
```

Pdf1 - EICHERMOT.pdf

In [15]:
```
documentPage = "./Rec_Task/EICHERMOT.pdf"
templatePage = "EICHERMOT.tabula-template.json"


dfs=tabula.read_pdf_with_template(documentPage, templatePage, stream=True)
```

In [39]:
```
tables = []
for x in dfs:
    dfObj = pd.DataFrame(x)
    tables.append(dfObj)
    display(dfObj)
```

	No.	NaN	Member	meetings	meetings
0	No.	NaN	Member	meetings	meetings
1	NaN	NaN	NaN	held	attended
2	1.	Mr S. Sandilya	Chairman	2	2
3	2.	Mr Siddhartha Lal	Member	2	2
4	3.	Mr Prateek Jalan	Member	2	2

	Sl.	Name	Chairman/
0	No.	NaN	Member
1	1.	Mr Siddhartha Lal (Managing Director &	Chairman
2	NaN	Chief Executive Officer)	NaN
3	2.	Mr S. Sandilya (Chairman and Non-Executive	Member
4	NaN	Independent Director)	NaN
5	3.	Mr Lalit Malik (Chief Financial Officer)	Member

	Unnamed: 0	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	(Rs. in Crores)
0	Name of the Director	Salary	Commission	Perquisites	NaN	Service Contract	NaN
1	NaN	NaN	NaN	NaN	Tenure	Notice period	NaN
2	Mr Siddhartha Lal	3.29	4.20	1.71	5 years (Members at their	3 months' notice or salary in	NaN
3	Managing Director	NaN	NaN	NaN	AGM held on June 18, 2016,	lieu of notice for 3 months or for	NaN
4	NaN	NaN	NaN	NaN	approved re-appointment of	such period which falls short of	NaN
5	NaN	NaN	NaN	NaN	Mr Siddhartha Lal as Managing	3 months	NaN
6	NaN	NaN	NaN	NaN	Director w.e.f. May 1, 2016 up to	NaN	NaN
7	NaN	NaN	NaN	NaN	April 30, 2021)	NaN	NaN

Special processing required for the tables(such as below example)

Table1:

In [40]:
```
display(tables[0])
```

	No.	NaN	Member	meetings	meetings
0	No.	NaN	Member	meetings	meetings
1	NaN	NaN	NaN	held	attended
2	1.	Mr S. Sandilya	Chairman	2	2
3	2.	Mr Siddhartha Lal	Member	2	2
4	3.	Mr Prateek Jalan	Member	2	2

In [57]:
```
df = tables[0]
df = df.iloc[1:]
df.columns = np.concatenate([df.columns[:3],df.iloc[0, 3:]])
df = df.iloc[1:]
df = df.reset_index()
del df['index']
```

In [63]:
```
display(df)
```

	No.	NaN	Member	held	attended
0	1.	Mr S. Sandilya	Chairman	2	2
1	2.	Mr Siddhartha Lal	Member	2	2
2	3.	Mr Prateek Jalan	Member	2	2

Pdf2 -File1.pdf

In [83]:
```
documentPage = "./Rec_Task/d9f8e6d9-660b-4505-86f9-952e45ca6da0.pdf"
templatePage = "./page2.json"
dfs1=tabula.read_pdf_with_template(documentPage, templatePage, stream=True)
```

```
Got stderr: Dec 22, 2020 7:13:18 PM org.apache.pdfbox.pdmodel.font.FileSystemFontProvider loadDiskCache
WARNING: New fonts found, font cache will be re-built
Dec 22, 2020 7:13:18 PM org.apache.pdfbox.pdmodel.font.FileSystemFontProvider <init>
WARNING: Building on-disk font cache, this may take a while
Dec 22, 2020 7:13:21 PM org.apache.pdfbox.pdmodel.font.FileSystemFontProvider <init>
WARNING: Finished building on-disk font cache, found 2602 fonts
```

In [84]:
```
tables1 = []
for x in dfs1:
    dfObj = pd.DataFrame(x)
    tables1.append(dfObj)
    display(dfObj)
```

	Date	Name of the analyst/investor	Type	Location
0	April 4, 2018	Motilal Oswal Asset Management	One-on-One	Mumbai
1	NaN	Company Limited	meeting	NaN
2	NaN	Credit Suisse	Voice call	-
3	NaN	I	NaN	NaN
4	April 5, 2018	Maybank Eng Securities India Private	One-on-One	Mumbai
5	NaN	Limited	meeting	NaN

In [ ]:
```

```

In [ ]:
```

```