In [1]:

```
from pytesseract import pytesseract
import os
import glob
import re
import pandas as pd
class OCR():
    def pan(self):
        data = glob.glob(r"C:\Users\User\Desktop\practice\117_Swapnil Badgujar_Mini project\PAN\*" + "*.jpeg")
        Pan_Data = {"PAN_IMAGE":[],"Pan_No":[],"Pan_Name":[],"Pan_DOB":[]}
        for pan in data:
            pytesseract.tesseract_cmd = r'C:\python\tesseract.exe'
            text = pytesseract.image_to_string(pan)
            text = re.sub("INCOME|TAX|DEPARTMENT|Signature|GOVT|OF|INDIA|Permanent|Account|Number|\W"," ",text)
            Pan_Data["PAN_IMAGE"].append((pan.split("\\")[-1]).replace(".jpeg",""))
            Pan_Data["Pan_No"].extend(re.findall("[A-Z]{5}[0-9]{4}[A-Z]{1}",text))
            Pan_Data["Pan_DOB"].extend(re.findall("\d{2}[/\s]\d{2}[/\s]\d{4}",text))
            Pan_Data["Pan_Name"].extend(re.findall("[A-Z]{1,10}\s{1,2}[A-Z]{1,10}\s{1,2}[A-Z]{1,15}",text))
        df = pd.DataFrame({ key:pd.Series(value) for key, value in Pan_Data.items() })
        df.to_csv('PAN_data2.csv',index = False)
        print(pd.read_csv("PAN_data2.csv"))
    def Aadhar(self):
        data = glob.glob(r"C:\Users\User\Desktop\practice\117_Swapnil Badgujar_Mini project\AADHAR\*" + "*.jpeg")
        Aadhar_Data = {"Aadhar_Image":[],"Aadhar_No":[],"Aadhar_DOB":[]}
        for Aadhar in data:
            pytesseract.tesseract_cmd = r'C:\python\tesseract.exe'
            text = pytesseract.image_to_string(Aadhar)
            text = re.sub("Name|Female|FEMAIL|MALE|Male|Gender|\n"," ",text)
            Aadhar_Data["Aadhar_Image"].append((Aadhar.split("\\")[-1]).replace(".jpeg",""))
            Aadhar_Data["Aadhar_No"].extend(re.findall("\d{4}\s\d{4}\s\d{4}",text))
            Aadhar_Data["Aadhar_DOB"].extend(re.findall("\d{2}[/\s-]\d{2}[/\s-]\d{4}",text))
            #Aadhar_Data["Aadhar_Name"].extend(re.findall("[A-Za-z]{3,10}\s[A-Za-z]{0,10}\s[A-Za-z]{0,10}",text))
        df = pd.DataFrame({ key:pd.Series(value) for key, value in Aadhar_Data.items() })
        df.to_csv('Aadhar_Data.csv',index = False)
        print(pd.read_csv("Aadhar_Data.csv"))

a = OCR()
a.pan()
a.Aadhar()
```

```
   PAN_IMAGE      Pan_No                        Pan_Name     Pan_DOB
0       pan1  BNZPM2501F         D MANIKANDAN  DURAISAMY  16 07 1986
1      pan10  AQSPL9772C               KUSUM LATA  DHANI  17 10 1992
2      pan11  GQBPK8700C                     GA AL AMAAB  04 05 1997
3       pan2  EJAPS0276M           MONIKA MAHADEV SHINDE  31 10 1992
4       pan3  BJDPP6011M  F  PREMSANKAR  VANAMAMALAIPERU  09 07 1986
5       pan4  ANRPM2537J              PRAMOD KUMAR MAHTO  03 04 1982
6       pan5  AQNPM7970Q          AASHISH MISHRA  MAHESH  17 09 1984
7       pan6  ANRPM2537J              PRAMOD KUMAR MAHTO  03 04 1982
8       pan7  BLQPK3045P            MANOJ KUMAR  NARURAM  01 12 1988
9       pan8  DUTPS3077K        SMITA PRAKASH SRIVASTAVA  05 02 1984
10      pan9  ANUPT5774F            PRAVESH PRASAD SINHA  10 11 1992
11       NaN         NaN                  MOHAMMD TA EEQ         NaN
   Aadhar_Image       Aadhar_No  Aadhar_DOB
0       Aadhar1  3425 0653 1151  28/05/2000
1      Aadhar10  6536 4848 7185  19/07/1995
2      Aadhar11  2312 5823 4114  25/08/1995
3      Aadhar12  5939 7553 9390  22/06/1983
4      Aadhar13  7109 5388 5107  23/10/2011
5       Aadhar2  8158 4542 1351  05-06-1965
6       Aadhar3  8158 4542 1351  05-06-1965
7       Aadhar4  5630 0841 0574  06/08/1999
8       Aadhar5  3425 0653 1151  28/05/2000
9       Aadhar6  2879 9185 1180  27/12 1088
10      Aadhar8  2114 5270 9955  11/08/1993
11      Aadhar9  2094 7051 9541  01/01/1959
```

In [2]:

```
from pytesseract import pytesseract
import os
import glob
import re
class OCR():
    def pan(self):
        data = glob.glob(r"C:\Users\User\Desktop\practice\117_Swapnil Badgujar_Mini project\PAN\*" + "*.jpeg")
        for pan in data:
            pytesseract.tesseract_cmd = r'C:\python\tesseract.exe'
            text = pytesseract.image_to_string(pan)
            text = re.sub("INCOME|TAX|DEPARTMENT|Signature|GOVT|OF|INDIA|Permanent|Account|Number|\W"," ",text)
            Pan_Data = {(pan.split("\\")[-1]).replace(".jpg","") : {"PAN_No":[],"PAN_DOB":[],"PAN_Name":[]}}
            Pan_Data[(pan.split("\\")[-1]).replace(".jpg","")]["PAN_No"].append(re.findall("[A-Z]{5}[0-9]{4}[A-Z]{1}
            Pan_Data[(pan.split("\\")[-1]).replace(".jpg","")]["PAN_DOB"].append(re.findall("\d{2}[/\s]\d{2}[/\s]\d{
            Pan_Data[(pan.split("\\")[-1]).replace(".jpg","")]["PAN_Name"].append(re.findall("[A-Z]{1,10}\s{1,2}[A-Z
            print(Pan_Data)
    def Aadhar(self):
        data = glob.glob(r"C:\Users\User\Desktop\practice\117_Swapnil Badgujar_Mini project\AADHAR\*" + "*.jpeg")
        for Aadhar in data:
            pytesseract.tesseract_cmd = r'C:\python\tesseract.exe'
            text = pytesseract.image_to_string(Aadhar)
            text = re.sub("Name|Female|FEMAIL|MALE|Male|Gender|Government of India|\n"," ",text)
            Aadhar_data ={(Aadhar.split("\\")[-1]).replace(".jpeg","") : {"Aadhar_No":[],"Aadhar_DOB":[],"Aadhar_Nam
            Aadhar_data[(Aadhar.split("\\")[-1]).replace(".jpeg","")]["Aadhar_No"].append(re.findall("\d{4}\s\d{4}\s
            Aadhar_data[(Aadhar.split("\\")[-1]).replace(".jpeg","")]["Aadhar_DOB"].append(re.findall("\d{2}[/\s-]\d
            #Aadhar_data[(Aadhar.split("\\")[-1]).replace(".jpeg","")]["Aadhar_Name"].append(re.findall("[A-Za-z]{3,
            print(Aadhar_data)

OCR().pan()
OCR().Aadhar()
```

```
{'pan1.jpeg': {'PAN_No': [['BNZPM2501F']], 'PAN_DOB': [['16 07 1986']], 'PAN_Name': [['D MANIKANDAN  DURAISAMY']]}}
{'pan10.jpeg': {'PAN_No': [['AQSPL9772C']], 'PAN_DOB': [['17 10 1992']], 'PAN_Name': [['KUSUM LATA  DHANI']]}}
{'pan11.jpeg': {'PAN_No': [['GQBPK8700C']], 'PAN_DOB': [['04 05 1997']], 'PAN_Name': [['GA AL AMAAB']]}}
{'pan2.jpeg': {'PAN_No': [['EJAPS0276M']], 'PAN_DOB': [['31 10 1992']], 'PAN_Name': [['MONIKA MAHADEV SHINDE']]}}
{'pan3.jpeg': {'PAN_No': [['BJDPP6011M']], 'PAN_DOB': [['09 07 1986']], 'PAN_Name': [['F  PREMSANKAR  VANAMAMALAIPER
U']]}}
{'pan4.jpeg': {'PAN_No': [['ANRPM2537J']], 'PAN_DOB': [['03 04 1982']], 'PAN_Name': [['PRAMOD KUMAR MAHTO']]}}
{'pan5.jpeg': {'PAN_No': [['AQNPM7970Q']], 'PAN_DOB': [['17 09 1984']], 'PAN_Name': [['AASHISH MISHRA  MAHESH']]}}
{'pan6.jpeg': {'PAN_No': [['ANRPM2537J']], 'PAN_DOB': [['03 04 1982']], 'PAN_Name': [['PRAMOD KUMAR MAHTO']]}}
{'pan7.jpeg': {'PAN_No': [['BLQPK3045P']], 'PAN_DOB': [['01 12 1988']], 'PAN_Name': [['MANOJ KUMAR  NARURAM']]}}
{'pan8.jpeg': {'PAN_No': [['DUTPS3077K']], 'PAN_DOB': [['05 02 1984']], 'PAN_Name': [['SMITA PRAKASH SRIVASTAVA', 'PRAV
ESH PRASAD SINHA']]}}
{'pan9.jpeg': {'PAN_No': [['ANUPT5774F']], 'PAN_DOB': [['10 11 1992']], 'PAN_Name': [['MOHAMMD TA EEQ']]}}
{'Aadhar1': {'Aadhar_No': [['3425 0653 1151']], 'Aadhar_DOB': [['28/05/2000']], 'Aadhar_Name': []}}
{'Aadhar10': {'Aadhar_No': [['6536 4848 7185']], 'Aadhar_DOB': [['19/07/1995']], 'Aadhar_Name': []}}
{'Aadhar11': {'Aadhar_No': [['2312 5823 4114']], 'Aadhar_DOB': [['25/08/1995']], 'Aadhar_Name': []}}
{'Aadhar12': {'Aadhar_No': [['5939 7553 9390']], 'Aadhar_DOB': [['22/06/1983']], 'Aadhar_Name': []}}
{'Aadhar13': {'Aadhar_No': [['7109 5388 5107']], 'Aadhar_DOB': [['23/10/2011']], 'Aadhar_Name': []}}
{'Aadhar2': {'Aadhar_No': [['8158 4542 1351']], 'Aadhar_DOB': [['05-06-1965']], 'Aadhar_Name': []}}
{'Aadhar3': {'Aadhar_No': [['8158 4542 1351']], 'Aadhar_DOB': [['05-06-1965']], 'Aadhar_Name': []}}
{'Aadhar4': {'Aadhar_No': [['5630 0841 0574']], 'Aadhar_DOB': [['06/08/1999']], 'Aadhar_Name': []}}
{'Aadhar5': {'Aadhar_No': [['3425 0653 1151']], 'Aadhar_DOB': [['28/05/2000']], 'Aadhar_Name': []}}
{'Aadhar6': {'Aadhar_No': [['2879 9185 1180']], 'Aadhar_DOB': [['27/12 1088']], 'Aadhar_Name': []}}
{'Aadhar8': {'Aadhar_No': [['2114 5270 9955']], 'Aadhar_DOB': [['11/08/1993']], 'Aadhar_Name': []}}
{'Aadhar9': {'Aadhar_No': [['2094 7051 9541']], 'Aadhar_DOB': [['01/01/1959']], 'Aadhar_Name': []}}
```