Extract Content from PDF

In [1]:
```
# !pip install mistral
```

In [2]:
```
# File URL you want to extract
FILE_URL  = "https://docs-ahmad.s3.us-east-1.amazonaws.com/Insurance.pdf"
```

In [3]:
```
import re
import json
import pandas as pd
from mistralai import Mistral

api_key = '5eVI4MbP8v5IttDlD7sYmhIpQrkuVlqA'
client = Mistral(api_key=api_key)

ocr_response = client.ocr.process(
    model="mistral-ocr-latest",
    document={
        "type": "document_url",
        "document_url": FILE_URL
    },
    include_image_base64=True
)
```

In [4]:
```
print(ocr_response.pages[0].markdown)
```

```
# Insurance Quotation

Number of lives per class and their premium costs table

| Class | Number of Lives | Premium Cost |
| :--: | :--: | :--: |
| VIP | 120 | 180,000 |
| A | 300 | 300,000 |
| B | 500 | 355,000 |
| C | 1000 | 500,000 |

عدد الأشخاص حسب الفئة والجنسية

| الفئة | الجنسية | عدد الأشخاص |
| :--: | :--: | :--: |
| 100 | أردني | VIP |
| 20 | غير أردني | VIP |
| 300 | أردني | A |
| 100 | غير أردني | A |
```

Document Understanding

In [5]:
```
from mistralai import Mistral

# Specify model
model = "pixtral-12b-2409"

# Initialize the Mistral client
client = Mistral(api_key=api_key)

text_query = """
    Give me Breakup of Census and Breakup of Rates tables in a usable manner.
    give me each table as dictoinary where column name is key and value is the values.
    So it can be ready to convert to a pandas dataframe.
    Please only return the data, I don't want anything additional.
    Please add <START_TABLE> tag and </START_TABLE> tag so it's easier for me to use. Please keep the languages
    The document probably contains English and arabic content so please extract them well"""

# Define the messages for the chat
messages = [
    {
        "role": "user",
        "content": [
            {
                "type": "text",
                "text": text_query
            },
            {
                "type": "document_url",
                "document_url": FILE_URL,
            }
        ]
```

```
    }
]

# Get the chat response
chat_response = client.chat.complete(
    model=model,
    messages=messages
)

# Print the content of the response
response = chat_response.choices[0].message.content
print(response)
```

```
```json
<START_TABLE>
{
    "Class": ["VIP", "A", "B", "C"],
    "Number of Lives": [120, 300, 500, 1000],
    "Premium Cost": ["180,000", "300,000", "355,000", "500,000"]
}
</START_TABLE>

<START_TABLE>
{
    "Number of People": [100, 20, 300, 100],
    "Nationality": ["أردني" ,"غير أردني" ,"أردني" ,"غير أردني"],
    "Class": ["VIP", "VIP", "A", "A"]
}
</START_TABLE>
```
```

In [6]:
```
text = response
# Update regex to match <START_TABLE> and </START_TABLE>
matches = re.findall(r'<START_TABLE>(.*?)</START_TABLE>', response, flags=re.DOTALL)

data_list = []
# Check if matches are found
if matches:
    for match in matches:
        # Clean up the match to make it valid JSON
        json_data = match.strip()
        try:
            # Parse the cleaned JSON string
            data = json.loads(json_data)
            data_list.append(data)
        except json.JSONDecodeError as e:
            print("Error parsing JSON:", e)
else:
    print("No tables found.")
```

In [7]:
```
pd.DataFrame(data_list[0])
```

Out[7]:

	Class	Number of Lives	Premium Cost
0	VIP	120	180,000
1	A	300	300,000
2	B	500	355,000
3	C	1000	500,000

In [8]:
```
pd.DataFrame(data_list[1])
```

Out[8]:

	Number of People	Nationality	Class
0	100	أردني	VIP
1	20	غير أردني	VIP
2	300	أردني	A
3	100	غير أردني	A

More info on https://docs.mistral.ai/capabilities/document/#document-understanding