## Importing Packages

In [ ]:

```
import pandas as pd
from pathlib import Path
import glob
from pypdf import PdfReader
from PyPDF2 import PdfMerger
import re
import csv
```

## Merge PDFs

In [ ]:

```
folder_path = Path(r'C:\Users\sular\OneDrive\1.Data Analysis & Engineering\Data Engineering\Data Enginee
```

In [ ]:

```
list(folder_path.rglob('*.pdf'))
```

Out[ ]:

```
[WindowsPath('C:/Users/sular/OneDrive/1.Data Analysis & Engineering/Data Engineering/Data Engineering P
rojects/LSG-PDF-Manifests-Extraction-Python/Data/Run Manifest (1).pdf'),
 WindowsPath('C:/Users/sular/OneDrive/1.Data Analysis & Engineering/Data Engineering/Data Engineering P
rojects/LSG-PDF-Manifests-Extraction-Python/Data/Run Manifest.pdf')]
```

In [ ]:

```
# Create an instance of PdfMerger() class
merger = PdfMerger()
```

In [ ]:

```
for file_name in folder_path.rglob('*'):
    # print(file_name)
    merger.append(file_name)

merger.write("Final.pdf")
merger.close()
```

In [ ]:

```
file_path = 'Final.pdf'
```

## Read PDF

In [ ]:

```
reader = PdfReader(file_path)
```

In [ ]:

```
full_text = []
for p in range(len(reader.pages)):
    page = reader.pages[p]
    # print(page.extract_text())
    lsg_text = page.extract_text()
    full_text.append(lsg_text)

# print(full_text)
```

In [ ]:

```
type(full_text)
```

Out[ ]:

```
list
```

In [ ]:

```
# converting list items to string
text_string = ''.join(full_text)
```

In [ ]:

```
type(text_string)
```

Out[ ]:

```
str
```

## Text Extraction

In [ ]:

```
# Use of regex
pattern = r'\b\d{7}\b.*\b\d{4}\b'
matches = re.findall(pattern, text_string)

print(matches)
```

```
['1080429 2000', '1080736 1000', '1079677 2000', '1079960 1000', '1079599 2000', '1079906 1000', '1080772
1000', '1079951 2000', '1080613 1000', '1079686 2000', '1080486 1000', '1080673 1000', '1079337 1000', '1
080748 1000', '1080595 1000', '1080610 1000', '1080386 1020', '1080625 1000', '1081013 1010', '1080628 10
00', '1079695 1000', '1080568 1000', '1080457 1000', '1080782 1000', '1080463 1000', '1079858 1000', '107
9885 1000', '1080778 1000', '1079978 1000', '1080133 1000', '1080466 1000', '1081014 1010', '1079879 100
0', '1081011 0010', '1080616 1000', '1080563 1000', '1079972 1000', '1080727 1000', '1080784 2000', '1080
604 1000', '1080448 1000', '1080622 1000', '1079413 1000', '1080661 1000', '1080436 1000', '1080766 100
0', '1080500 1000', '1080718 1000', '1079384 1000', '1080712 1000', '1080579 1000', '1080637 1000', '1080
539 1000', '1081040 1010', '1080492 1000', '1081042 1000', '1080262 1010', '1080497 1010', '1080320 201
0', '1080064 1000', '1080065 0010', '1080846 0020', '1080850 1010', '1080069 1000', '1080070 1000', '1080
769 1000', '1080560 1000', '1080475 1000', '1080775 1000', '1080795 1000', '1080607 1000', '1080676 200
0', '1080652 1000', '1079918 1000', '1079369 1000', '1080396 1000', '1080849 1010', '1080619 3000', '1080
670 1000', '1080679 1000', '1080763 1000', '1080478 1000', '1080757 1000', '1080700 1000', '1080489 100
0', '1080797 1000']
```

```
In [ ]: first_numbers = []
        second_numbers = []

        for item in matches:
            first, second = item.split()
            # first, second = map(int, item.split())
            first_numbers.append(first)
            second_numbers.append(second)

        print("First numbers:", first_numbers)
        print("Second numbers:", second_numbers)



        def digit_sum(number):
            return sum(int(digit) for digit in number)

        total_sums = [digit_sum(num) for num in second_numbers]

        #print("Original list:", second_numbers)
        print("Total sums list:", total_sums)
```

```
First numbers: ['1080429', '1080736', '1079677', '1079960', '1079599', '1079906', '1080772', '1079951',
'1080613', '1079686', '1080486', '1080673', '1079337', '1080748', '1080595', '1080610', '1080386', '10806
25', '1081013', '1080628', '1079695', '1080568', '1080457', '1080782', '1080463', '1079858', '1079885',
'1080778', '1079978', '1080133', '1080466', '1081014', '1079879', '1081011', '1080616', '1080563', '10799
72', '1080727', '1080784', '1080604', '1080448', '1080622', '1079413', '1080661', '1080436', '1080766',
'1080500', '1080718', '1079384', '1080712', '1080579', '1080637', '1080539', '1081040', '1080492', '10810
42', '1080262', '1080497', '1080320', '1080064', '1080065', '1080846', '1080850', '1080069', '1080070',
'1080769', '1080560', '1080475', '1080775', '1080795', '1080607', '1080676', '1080652', '1079918', '10793
69', '1080396', '1080849', '1080619', '1080670', '1080679', '1080763', '1080478', '1080757', '1080700',
'1080489', '1080797']
Second numbers: ['2000', '1000', '2000', '1000', '2000', '1000', '1000', '2000', '1000', '2000', '1000',
'1000', '1000', '1000', '1000', '1000', '1020', '1000', '1010', '1000', '1000', '1000', '1000', '1000',
'1000', '1000', '1000', '1000', '1000', '1000', '1000', '1010', '1000', '0010', '1000', '1000', '1000',
'1000', '2000', '1000', '1000', '1000', '1000', '1000', '1000', '1000', '1000', '1000', '1000', '1000',
'1000', '1000', '1000', '1010', '1000', '1000', '1010', '1010', '2010', '1000', '0010', '0020', '1010',
'1000', '1000', '1000', '1000', '1000', '1000', '1000', '1000', '2000', '1000', '1000', '1000', '1000',
'1010', '3000', '1000', '1000', '1000', '1000', '1000', '1000', '1000', '1000']
Total sums list: [2, 1, 2, 1, 2, 1, 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 3, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 2, 2, 3, 1, 1, 2, 2, 1,
1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 2, 3, 1, 1, 1, 1, 1, 1, 1, 1]
```

## Exporting csv

```
In [ ]: csv_file = open('LSG_Quantities.csv', 'w', newline='')
        csv_writer = csv.writer(csv_file)
        csv_writer.writerow(['Sales Purchase Order Number', 'Quantities', 'Total'])

        for first, second, third in zip(first_numbers, second_numbers, total_sums):
            csv_writer.writerow([first, second, third])

        csv_file.close()

        print("CSV file 'LSG_Quantities.csv' created.")
```

```
CSV file 'LSG_Quantities.csv' created.
```