Create dataframe from reading differnt formats

Ref: https://sparkbyexamples.com/pandas/

```
1 #import libraries
2 import pandas as pd
```

Mount my Gdirve

```
1 from google.colab import drive
2 ROOT = "/content/GDrive"
3 drive.mount(ROOT)
```

```
Mounted at /content/GDrive
```

Excel format

```
1 # Read Excel file
2 excel_path = "/content/GDrive/MyDrive/COS3302/week3/dada_files/courses.XLSX"
3 df = pd.read_excel(excel_path)
4 print(df)
```

```
  Courses     Fee Duration   Discount
0   Spark  25000   50 Days       2000
1  Pandas  20000   35 Days       1000
2    Java  15000       NaN        800
3  Python  15000   30 Days        500
4     PHP  18000   30 Days        800
```

```
1 # Read excel by considering first row as data
2 columns = ["courses","course_fee","course_duration","course_discount"]
3 df2 = pd.read_excel(excel_path, header=0, names = columns)
4 df2
```

	courses	course_fee	course_duration	course_discount
0	Spark	25000	50 Days	2000
1	Pandas	20000	35 Days	1000
2	Java	15000	NaN	800
3	Python	15000	30 Days	500
4	PHP	18000	30 Days	800

```
1 # Read excel by setting column as index
2 df2 = pd.read_excel(excel_path, index_col=0)
3 print(df2)
```

```
             Fee Duration   Discount
Courses
Spark      25000   50 Days       2000
Pandas     20000   35 Days       1000
Java       15000       NaN        800
Python     15000   30 Days        500
PHP        18000   30 Days        800
```

```
1 # Read specific excel sheet
2 df = pd.read_excel(excel_path, sheet_name="Sheet1")
```

```
print(df)
```

```
  Courses     Fee Duration   Discount
0    Spark  25000   50 Days       2000
1   Pandas  20000   35 Days       1000
2     Java  15000        NaN       800
3   Python  15000   30 Days        500
4      PHP  18000   30 Days        800
```

```
# Read Multiple sheets
dict_df = pd.read_excel(excel_path,
                   sheet_name=["Sheet1","Sheet2"])

# Get DataFrame from Dict
course_df = dict_df.get("Sheet1")
course_updated_df = dict_df.get("Sheet2")

# Print DataFrame's
print(course_df)
print(course_updated_df)
```

```
  Courses     Fee Duration   Discount
0    Spark  25000   50 Days       2000
1   Pandas  20000   35 Days       1000
2     Java  15000        NaN       800
3   Python  15000   30 Days        500
4      PHP  18000   30 Days        800
  Courses     Fee Duration   Discount
0    Spark  25000   50 Days       2000
1   Pandas  20000   35 Days       1000
2     Java  15000   30 Days        700
3   Python  15000   30 Days        500
4      PHP  18000   30 Days        800
```

```
# Read excel by skipping columns
df2 = pd.read_excel(excel_path, usecols=[0,2])
print(df2)
```

```
  Courses  Duration
0    Spark   50 Days
1   Pandas   35 Days
2     Java       NaN
3   Python   30 Days
4      PHP   30 Days
```

```
# Skip columns by range
df2 = pd.read_excel(excel_path, usecols='B:D')
print(df2)
```

```
     Fee Duration   Discount
0  25000   50 Days       2000
1  20000   35 Days       1000
2  15000       NaN        800
3  15000   30 Days        500
4  18000   30 Days        800
```

```
# Read excel file by skipping rows
df2 = pd.read_excel(excel_path, skiprows=2)
print(df2)
```

```
    Pandas  20000  35 Days  1000
0     Java  15000      NaN   800
1   Python  15000  30 Days   500
2      PHP  18000  30 Days   800
```

```
# Using skiprows to skip rows
df2 = pd.read_excel(excel_path,
                   skiprows=[1,3])
print(df2)
```

```
  Courses     Fee Duration   Discount
0   Pandas  20000   35 Days       1000
1   Python  15000   30 Days        500
2      PHP  18000   30 Days        800
```

```
1 # Using skiprows with lambda
2 df2 = pd.read_excel(excel_path,
3                     skiprows=lambda x: x in [1,3])
4 print(df2)
```

```
  Courses     Fee Duration   Discount
0   Pandas  20000   35 Days       1000
1   Python  15000   30 Days        500
2      PHP  18000   30 Days        800
```

Json format

```
1 # Read json from String
2 json_str = '{"Courses":{"r1":"Spark"},"Fee":{"r1":"25000"},"Duration":{"r1":"50 Days"}}'
3 df = pd.read_json(json_str)
4 print(df)
```

```
   Courses    Fee Duration
r1   Spark  25000  50 Days
```

```
1 # Read json from String
2 json_str = '[{"Courses":"Spark","Fee":"25000","Duration":"50 Days","Discount":"2000"}]'
3 df = pd.read_json(json_str, orient='records')
4 print(df)
```

```
  Courses    Fee Duration  Discount
0   Spark  25000  50 Days      2000
```

```
1 json_file = "/content/GDrive/MyDrive/COS3302/week3/dada_files/courses_data.json"
2 df = pd.read_json(json_file)
3 print(df)
```

```
  Courses    Fee Duration
0   Spark  25000  50 Days
1  Pandas  20000  35 Days
2    Java  15000
```

```
1 # Read JSON file with records orient
2 df = pd.read_json("/content/GDrive/MyDrive/COS3302/week3/dada_files/courses.json", orient='records')
3 print(df)
```

```
  Courses    Fee Duration  Discount
0   Spark  25000  50 Days      2000
1  Pandas  20000  35 Days      1000
2    Java  15000                800
```

CSV format (exercise)

✓ 0s completed at 4:55 PM