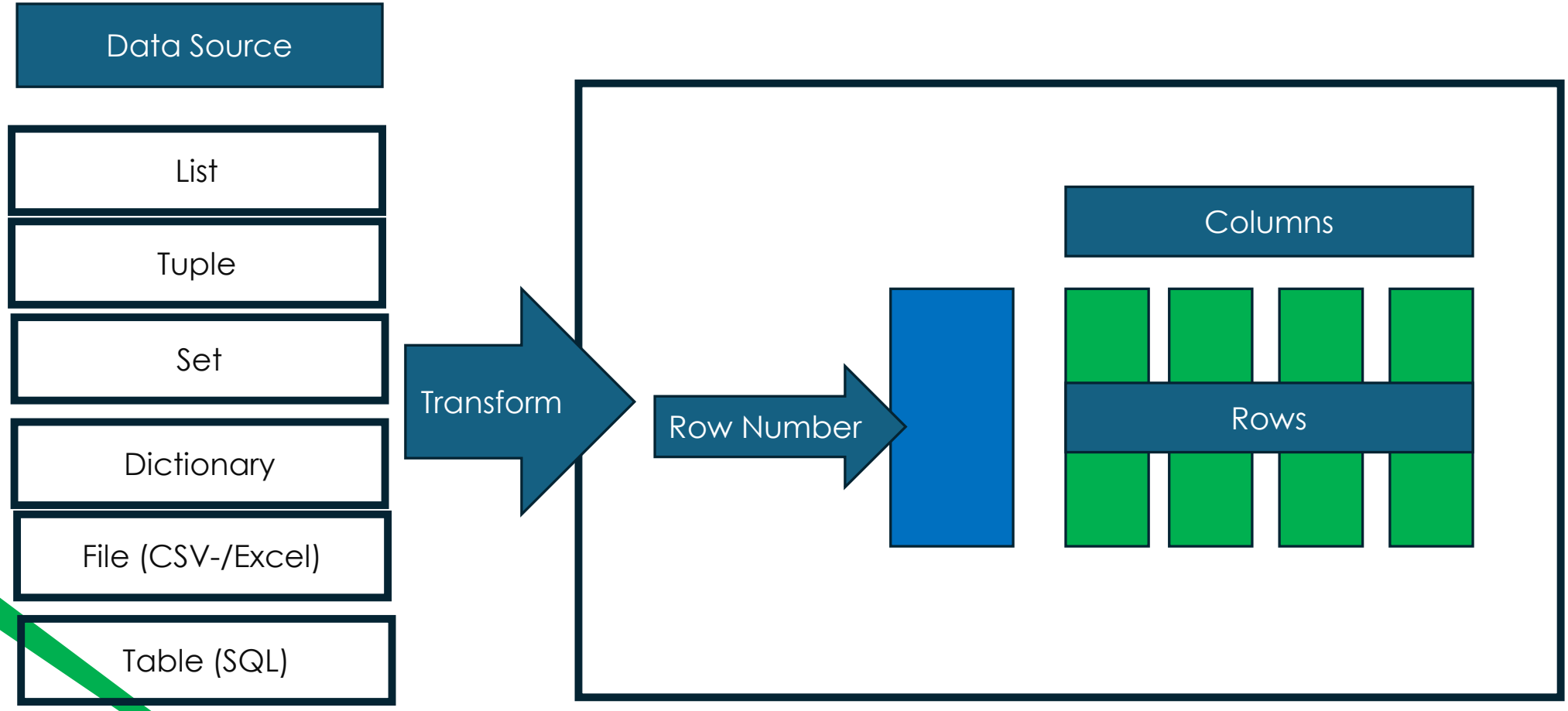
Agenda for Day #5

04/22/2024

- Pandas
- EDA (Exploratory Data Analysis)
- Data Visualization & SQL

DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore



DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore

```
data = [['Alex',10],['Bob',12],['Clarke',13]]
```

```
df = pd.DataFrame(data)
```

	0	1
0	Alex	10
1	Bob	12
2	Clarke	13

Column

In the list, there is no column values

```
df = pd.DataFrame(data,columns=['Customer Name','Customer Age'])
```

Data Frame

Data Source (List)

In the Data Source (List) there is no column name and we wanted to give the column name

	Customer Name	Customer Age
0	Alex	10
1	Bob	12
2	Clarke	13

```
data = {'Name':['Tom', 'Jack', 'Steve', 'Ricky'],'Age':[28,34,29,11]}
```

Key

The values are given as a list

Key

The values are given as a list

	Name	Age
0	Tom	28
1	Jack	34
2	Steve	29
3	Ricky	11

```
data = [
            {'a': 1, 'b': 2, 'c': 3},
            {'a': 5, 'b': 10, 'd': 20}
        ]
```

List

List of Dictionary

	a	b	c	d
0	1	2	3.0	NaN
1	5	10	NaN	20.0

There is no value for this column and it puts NaN

There is no value for this column and it puts NaN

In the Data Frame, we can add and remove columns in the data frame based on the data source.

In the Data Source, if column does not exist, you can add a column in the data frame without impacting the data source.

CID	PID	OID	ODATE	Quantity	Unit Price
C1	P1	11			
C1	P2	11			
C1	P3	11			

Use this as a data source in the data frame and calculate the column amount = Quantity * Unit Price.

DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore

Add Column

Remove Column

Change the Column Orders

```
x = pd.DataFrame()
```

```
x = pd.Series()
```

Rows , Column, and Index

```
Data_Source = {'KEY1' : 0., 'KEY2' : 1., 'KEY3' : 2.}
```

```
a = pd.Series(Data_Source)
```

```
KEY1   0.0
KEY2   1.0
KEY3   2.0
```

One Column and Multiple Rows

```
d = {'STOCK-INDEX-KEY-1' : pd.Series([1, 2, 3], index=['INDEX-1', 'INDEX-2', 'INDEX-3']),
     'STOCK-INDEX-KEY-2' : pd.Series([1, 2, 3, 4], index=['INDEX-1', 'INDEX-2', 'INDEX-3', 'INDEX-4'])}
```

A dictionary transformed in to Data Frame and Series

Series

```
ROW-1    4244444.0
ROW-2     385555.0
ROW-3          NaN
ROW-4          NaN
day3      396666.0
dtype: float64
```

	calories	duration
0	420	50
1	380	40
2	390	45

Data Frame

```
print(df.loc[2])
```

Locate Function, To Locate a Rows and Columns

Row Index

```
print(df.loc[[0, 1]])
```

Locate Row “0”

Locate Row “1”

Locating Multiple Rows and it’s column Values

```python
import pandas as pd
df = pd.read_csv('/content/sample_data/UNIT3LVADSAIEMPDATAv1.csv')
```

In Pandas a function to read a CSV File

DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore

```python
import csv
with open('/content/sample_data/UNIT3LVADSAIEMPDATAv1.csv', 'r') as file:
```

An alternate option to read a CSV File, there is library called CSV use this

DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore

```python
import pandas as pd
df = pd.read_csv('/content/sample_data/UNIT3LVADSAIEMPDATAv1.csv', na_values=['NA',
'NULL', 'Missing'])
```

This Handles NULL VALUES

This is the value I wanted to Substitute for NA or NULL

NA or NULL

DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore

```python
import pandas as pd
df = pd.read_json('/content/sample_data/UNIT3LVADSAI-WEBUSER.json')
print(df.to_string())
```

Read JASON File

```
print(df.shape)
```

Number of rows and columns

```
vAR_df = df.aggregate(["sum"])
vAR_df = df.aggregate(["min"])
vAR_df = df.aggregate(["max"])
```

This is similar to your SQL Aggreate Functions

```
print(df.mean(skipna = True))
```

CALC MEAN

Skip NULL or NA or NaN Values

DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore

```
vAR_sum = df[['COURSES-FEE','COURSES-DSICOUNT']].aggregate('min')
print(vAR_sum)
```

Only for Selected Columns in the Data Frame

AGG Function

Find MIN VALUE

DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore

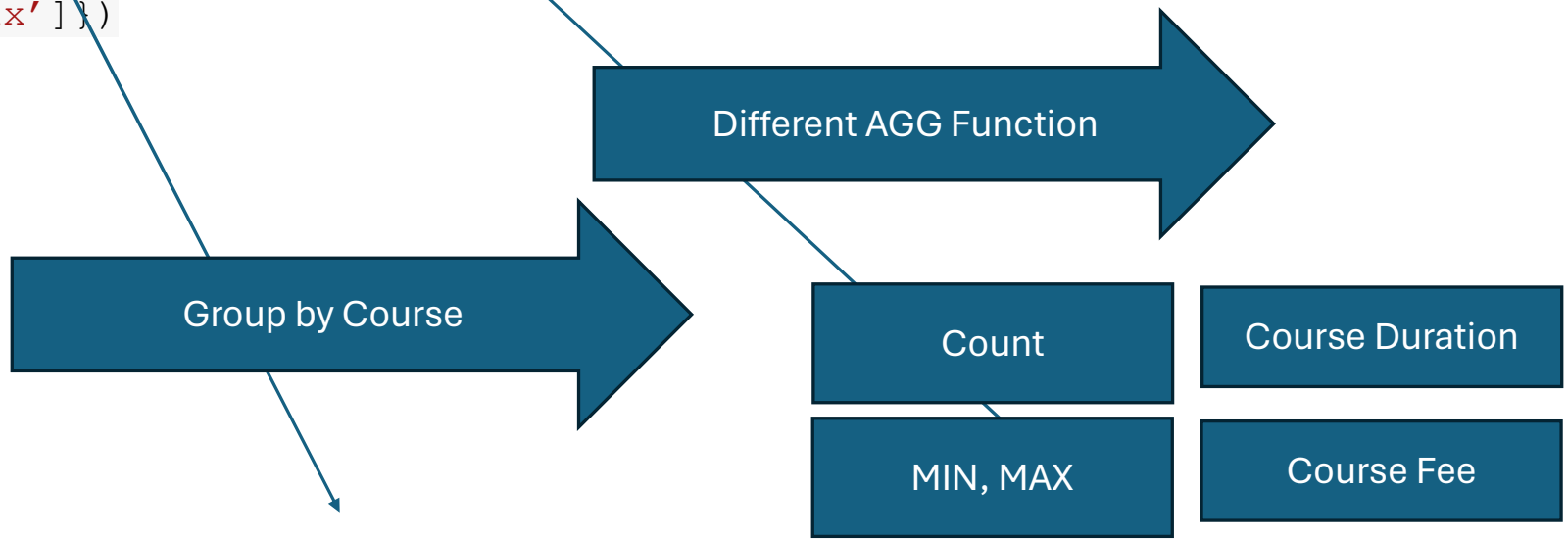
```
avg = df.apply(custom_mean, axis=1)
print(avg)
```

```
df = pd.DataFrame(vAR_Training)

# Groupby multiple columns & multiple aggregations

result = df.groupby('COURSES').aggregate({'COURSES-DURATION':'count','COURSES-FEE':['min','max']})

print(result)
```



Problem Statement

Analytics

Problem Statement: Distribution of customers between Male and female

Collect Data
(1000)

EDA
(Male : 800
Female : 200

Find out whether the given data has equal number of male and female distribution

DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore

EDA
(Male : 800
Female : 200

Objective of EDA

- Missing Values
- Null Values
- Distribution of Values
- Outliers
- Duplicate
- etc

DeepSphere.AI
Enterprise AI and IIoT for Analytics
USA | Singapore

Data Collection

EDA
(Male : 800
Female : 200

Data Preparation

	EMPLOYEE_ID	SALARY	DEPARTMENT_ID
count	75.000000	75.000000	75.000000
mean	135.946667	5073.546667	55.066667
std	27.429957	4060.602479	20.754246
min	100.000000	2100.000000	10.000000
25%	118.500000	2600.000000	50.000000
50%	136.000000	3200.000000	50.000000
75%	139.000000	6500.000000	50.000000
max	206.000000	24000.000000	110.000000

```
df.describe()
```

```
df.duplicated().sum()
```

Count of duplicate records

All columns in the DF should have the same values

DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore

Duplicate

EID	ENAME	SAL
101	JOTHI	100K
101	JOTHI	100K
101	JOTHI	100K
101	JOTHI	100K
101	JOTHI	100K
102	JOTHI	100K
103	JOTHI	100K

```
df['MANAGER_ID'].unique()
```

```
df.isnull().sum()
```

Count NULL Values in each column in the data frame

DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore

```
df[df['DEPARTMENT_ID']==10].head()
```

Like a where in the SQL

Look for this column in the data frame and see if the value is 10

DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore

DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore

DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore

DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore

DeepSphere.AI
Enterprise AI and IIoT for Analytics

USA | Singapore