DSBDA

* pandas - is a python library used in data analysis. It provides various data structures for manipulating & analyzing data.
  → import pandas as pd
  → df = pd.read_csv('path')

* numpy - (Numerical Python)
numpy consists of multi-dimensional array objects & collections to process array. It is used for working with array. It is used to perform math operations (linear algebra, matrix, fourier transform)
  → import numpy as np

* Dataframe -
It is 2D tabular data structure with labeled axis (rows & columns)

* Transposing -
It is used to permute the dimensions of array.

* shape & Reshape -
Shape - To give tuple of array dimensions
Reshape - To change dimensions of array into new shape.

* melt() -
converts data from wide format to long format.

* Data cleaning -
It means fixing bad data.
bad data - empty cells,
data in wrong format
wrong data
Duplicates

(a) Remove empty cells
i) new_df = df.dropna()
ii) df.Fillna(0)

(b) Data in wrong format
i) pd.to_datetime(df['Date'])

(c) wrong data
df.loc[row_no, col_name] = value

(d) Duplicates
df.duplicated() # True/False
df.drop_duplicates(inplace=True)

* Data integration -
It is a data processing technique that involves combining data from multiple data sources into coherent data store & provide unified view of data.

a) Read from csv file
import pandas as pd
pd.read_csv('heart.csv')

b) Read from google sheet
import pandas as pd
new_google_sheet_url = '<url>'
df = pd.read_csv(new_google_sheet_url)

c) Read Data from Database

```
import pandas as pd
import sqlite3
#Read via SQLite database
con = sqlite3.connect("your-db-link")
#Read table via select stmt
player = pd.read_sql_query("SELECT * from Table", con)
#close the connection
con.close()
```

* Data transformation

The process of converting raw data into a format or structure that would be more suitable to make model. It transform values.

a) # Multiply each value by 10 in array.

```
import pandas as pd
import numpy as np
df = pd.DataFrame(np.array([[1,2,3],
                            [4,5,6],
                            [7,8,9]]),
                  columns=['a','b','c'])
df.transform(func = lambda x: x*10)
```

b)

```
import pandas as pd
df.groupby('C')['A'].mean()
mean = df.groupby('C')["A"].mean().
       rename("N").reset_index()
df_1 = df.merge(mean)
df['N3'] = df.groupby(['C'])['A'].transform
           ('mean')
```

* matplotlib -
It is a plotting library used to plot graphs.
matplotlib.pyplot - provides functns.

* seaborn -
It is a library based on matplotlib which provides interface to draw attractive & informative statistical graph.

* sklearn
It is a tool used to build statiscal models including classification, regression, clustering etc.

1. Data Preparation

*enlivo
Date: __/__/____
Page: _____

* Dataset - heart.csv
* Operations - 1) Find shape of data
2) Find missing values
3) Find data type of each col
4) Find out zero's
5) Find mean age of patient

→

```
import pandas as pd
df = pd.read_csv('heart.csv')
1) df.shape
2) df.isnull()
3) df.dtypes
4) df.isnull().sum()
5) df['Age'].mean()
```

4/5. Data Preparation

* dataset - facebook & temperature
* operation - 1) creata data subset
  2) Mege data
  3) Sort data
  4) Transposing data
  5) Shape & Reshape data

→

```
import pandas pd
df = pd.read_csv(' ')
```

1)
```
# 1st 5 rows
  df.head()
# Last 5 rows
  df.tail()
# With one column
  new_df = df['like']
# With multiple columns
  new_df = df[['like', 'comment']]
# condition
  new_df = df[df['like']>35]
```

2)
```
a1 = df[['comment', 'like']]
a2 = df[['like', 'share']]
merge = pd.merge(a1, a2)
```

3)
```
df.sort_values('like', ascending=True)
```

4)
```
df.transpose
```

5)
```
df.shape
r1 = df.pivot_table(index=['Type', 'category'],
                    values=['comment'])
```

```
r2 = df.pivot_table(index=['Type','category'],
                    values=['comment',
                            'like'])
r3 = df.pivot_table(index=['Type','category'],
                    columns=['',''],
                    values=['comment'])
pd.melt(df)
```

6/7. Data Preparation

* Dataset - heart.csv & airquality.csv
* Operation - 1) Data cleaning
  2) Data Integration
  3) Data Transformation
  4) Error correcting
  5) Data model building

- df.loc[row.no, col_name]=value ←
- df['sex'] = df['sex'].replace([0,1],['female','male'])

→

1) df.fillna(0)
   df1 = df.dropna()
   df2 = df1.drop_duplicates()

2) x = df2['Age'].mean()
   y = df2['Age'].median()
   z = df2['Age'].mode()[0]

3) df3 = df2.groupby('ca')['Age'].mean()
   df4 = df3.rename("NewAge").reset_index()
   df5 = pd.merge(df2, df4)

5) from sklearn import linear_model
   from sklearn.metrics import mean_squared_error

> model = linear_model.LinearRegression()
> model
> xx = df['Age']
  yy = df['chol']
> model.fit(df[['chol']], df.Age)
> model.coef
> model.intercept
> y_pred = model.predict(df[['chd']])
> y_pred >print('MSE:', mean_squared_error(yy, y_pred))

8 to 11. Graph

* Dataset: forest fire, temperature, air quality, heart dataset.
* Operation :-
  Use matplotlib & seaborn
  1) Pie chart
  2) Bar graph
  3) Box plot
  4) line graph
  5) scatter plot graph

→

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
df = pd.read_csv(" ")
```

1)
```
plt.title("Pie Chart")
new_df = df.head()
x = new_df.Age
y = new_df.Thal
plt.pie(x, labels = y)
plt.show()
```

2)
```
plt.titale("Bar Graph")
plt.xlabel("AGE")
plt.ylabel("CHOL")
plt.bar(x, y)
```

3)
```
data = df.Chol
plt.title("Box Plot")
plt.xlabel("Chol")
plt.boxplot(data)
```

4)
```
x = df.Sex
y = df.Slope
plt.title("Line Graph")
plt.xlabel("SEX")
plt.ylabel("SLOPE")
plt.plot(x,y)
```

5)
```
plt.title("Scatter Graph")
plt.xlabel("AGE")
plt.ylabel("CHOL")
plt.scatter(x,y)
```

6)
```
sns.boxplot(x="Slope", y="Chol", data=df)
sns.scatterplot(x="Age", y="Chol", data=df)
sns.lineplot(data=df["Age"])
```