Data collection,Modelling and Compilation

Data collection

Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes. Data collection is a research component in all study fields, including physical and social sciences, humanities, and business. While methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same. The goal for all data collection is to capture quality evidence that allows analysis to lead to the formulation of convincing and credible answers to the questions that have been posed. Data collection and validation consists of four steps when it involves taking a census and seven steps when it involves sampling

Data collection from Datasets from csv files and Excel files

Creating a dataframe

In [4]:
```
my_dict = {'name': ['a','b','c','d','e','f','g'],'age' :[20,27,35,55,18,21,35],'designation': ["VP","CEO","CFO","VP","VP","CEO","MD"]}
import pandas as pd
import numpy as np
df=pd.DataFrame(my_dict)
df
```

Out[4]:

	name	age	designation
0	a	20	VP
1	b	27	CEO
2	c	35	CFO
3	d	55	VP
4	e	18	VP
5	f	21	CEO
6	g	35	MD

Saving a dataframe to a CSV file

In [6]:
```
df_to_csv('csv_fds')
df
```

Out[6]:

	name	age	designation
0	a	20	VP
1	b	27	CEO
2	c	35	CFO
3	d	55	VP
4	e	18	VP
5	f	21	CEO
6	g	35	MD

Loading CSV file as a dataframe

In [7]:
```
df_to_csv('csv_fds',index=False)
df_csv=pd.read_csv('csv_fds')
df_csv
```

Out[7]:

	name	age	designation
0	a	20	VP
1	b	27	CEO
2	c	35	CFO
3	d	55	VP
4	e	18	VP
5	f	21	CEO
6	g	35	MD

Loading data from a CSV file

In [12]:
```
import pandas as pd
location = "D:\Datasdet\students.csv"
df = pd.read_csv(location, header=None)
df.head()
```

Out[12]:

	0	1	2	3	4	5	6	7	8
0	id	first_name	last_name	date_of_birth	roll-no	Grades	BS	MS	PHD
1	1	John	Doe	Jan-00	11	75	1	1	2
2	2	Jane	Smith	May-01	15	80	2	1	1
3	3	Sarah	Thomas	Sep-02	20	90	1	1	1
4	4	Frank	Brown	Apr-02	21	97	2	1	1

Creating a dataframe using multiple lists

In [13]:
```
import pandas as pd
names = ['Bunny','Rohan','Mary','Raj','Sam']
grades = [78,74,75,88,90]
bsdegrees = [1,0,1,1,0]
msdegrees = [2,1,2,1,1]
phddegrees = [0,1,0,1,0]
degrees = zip(names,grades,bsdegrees,msdegrees,phddegrees)
columns = ['Names','Grades','BS','MS','PHD']
df = pd.DataFrame(data = degrees, columns=columns)
df
```

Out[13]:

	Names	Grades	BS	MS	PHD
0	Bunny	78	1	2	0
1	Rohan	74	0	1	1
2	Mary	75	1	2	0
3	Raj	88	1	1	1
4	Sam	90	0	1	0

Loading data from Excel files into dataframes

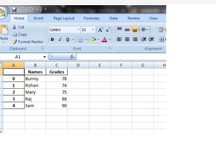
In [14]:
```
import pandas as pd
location = "D:\Datasdet\share.xlsx"
df = pd.read_excel(location)
df.columns = ['Roll no','first','last','sex','age','exer','hrs','grd','addr']
df.head()
```

Out[14]:

	Roll no	first	last	sex	age	exer	hrs	grd	addr
0	1	Chetan	Bhoir	M	22	3	10	75	Shelar
1	2	Simran	Patil	F	21	2	5	90	Aarogavn
2	3	Sagar	Gupta	M	20	1	8	81	Shivaji Chowk
3	4	Rahul	Gupta	M	23	2	9	95	Anjorphata
4	5	Dilip	Deo	M	25	3	5	90	Bhiwandi

Saving a dataframe to a Excel file

In [15]:
```
import pandas as pd
names = ['Bunny','Rohan','Mary','Raj','Sam']
grades = [78,74,75,88,90]
gradelist = zip(names,grades)
df = pd.DataFrame(data = gradelist,columns=['Names','Grades'])
writer = pd.ExcelWriter('dataframe_XSL.xlsx', engine='xlsxwriter')
df.to_excel(writer, sheet_name='sheet1')
writer.save()
```

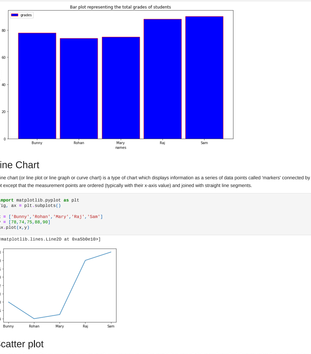


Bar Chart

A bar chart or bar graph is a chart or graph that represents categorical data with rectangular bars with heights or lengths proportional to the values they represent. These bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.

A bar chart can be created in Python by using matplotlib.pyplot.bar() or matplotlib.axes._subplots.AxesSubplot.bar() function.

In [18]:
```
import matplotlib.pyplot as plt

plt.figure(figsize = (12,7))

names = ['Bunny','Rohan','Mary','Raj','Sam']

grades = [78,74,75,88,90]

plt.bar(names, grades, width= 0.9, align='center',color='blue', edgecolor = 'red')

i = 1.0
j = 2000

for i in range(len(names)):
    plt.annotate(grades[i], (-0.1 + i, grades[i] + j))

plt.legend(labels = ['grades'])

plt.title("Bar plot representing the total grades of students")

plt.xlabel('names')
plt.ylabel('grades')

plt.savefig('1BarPlot.png')
```

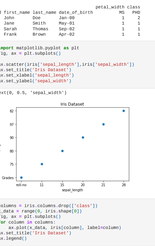


Line Chart

A line chart (or line plot or line graph or curve chart) is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. It is similar to a scatter plot except that the measurement points are ordered (typically with their x-axis value) and joined with straight line segments.

In [17]:
```
import matplotlib.pyplot as plt
fig, ax = plt.subplots()

x = ['Bunny','Rohan','Mary','Raj','Sam']
y = [78,74,75,88,90]
ax.plot(x,y)
```

Out[17]:



Scatter plot

A scatter plot (also called as scatterplot, scatter graph, scattergram or scatter graph) is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.

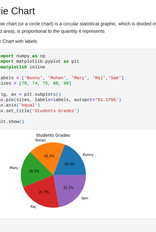
In [18]:
```
import pandas as pd
iris = pd.read_csv("D:\Datasdet\iris.csv", names=['sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'class'])
print(iris.head())
```

```
                                    sepal_length sepal_width petal_length  \
id first_name last_name date_of_birth    roll-no       Grades          BS
1  John       Doe       Jan-00             11           75           1
2  Jane       Smith     May-01             15           80           2
3  Sarah      Thomas    Sep-02             20           90           1
4  Frank      Brown     Apr-02             21           97           2

                                    petal_width class
id first_name last_name date_of_birth          MS   PHD
1  John       Doe       Jan-00                1     2
2  Jane       Smith     May-01                1     1
3  Sarah      Thomas    Sep-02                1     1
4  Frank      Brown     Apr-02                1     1
```

In [19]:
```
import matplotlib.pyplot as plt
fig, ax = plt.subplots()

ax.scatter(iris['sepal_length'],iris['sepal_width'])
ax.set_title('Iris Dataset')
ax.set_xlabel('sepal_length')
ax.set_ylabel('sepal_width')
```

Out[19]:

```
Text(0, 0.5, 'sepal_width')
```



In [20]:
```
columns = iris.columns.drop(['class'])
x_data = range(0, iris.shape[0])
fig, ax = plt.subplots()
for column in columns:
    ax.plot(iris.data, iris[column], label=column)
ax.set_title('Iris Dataset')
ax.legend()
```

Out[20]:

```
<matplotlib.legend.Legend at 0xa66021b>
```


In [21]:
```
fig, ax = plt.subplots()
ax.hist(iris['sepal_length'])
ax.set_title('Iris')
ax.set_xlabel('sepal_length')
ax.set_ylabel('Frequency')
```

Out[21]:

```
Text(0, 0.5, 'Frequency')
```


Pie Chart

A pie chart (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents

Pie Chart with labels

In [23]:
```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

labels = ['Bunny', 'Rohan','Mary', 'Raj','Sam']
sizes = [78, 74, 75, 88, 90]

fig, ax = plt.subplots()
ax.pie(sizes, labels=labels, autopct='%1.1f%%')
ax.axis('equal')
ax.set_title('Students Grades')

plt.show()
```


In [ ]:
```

```