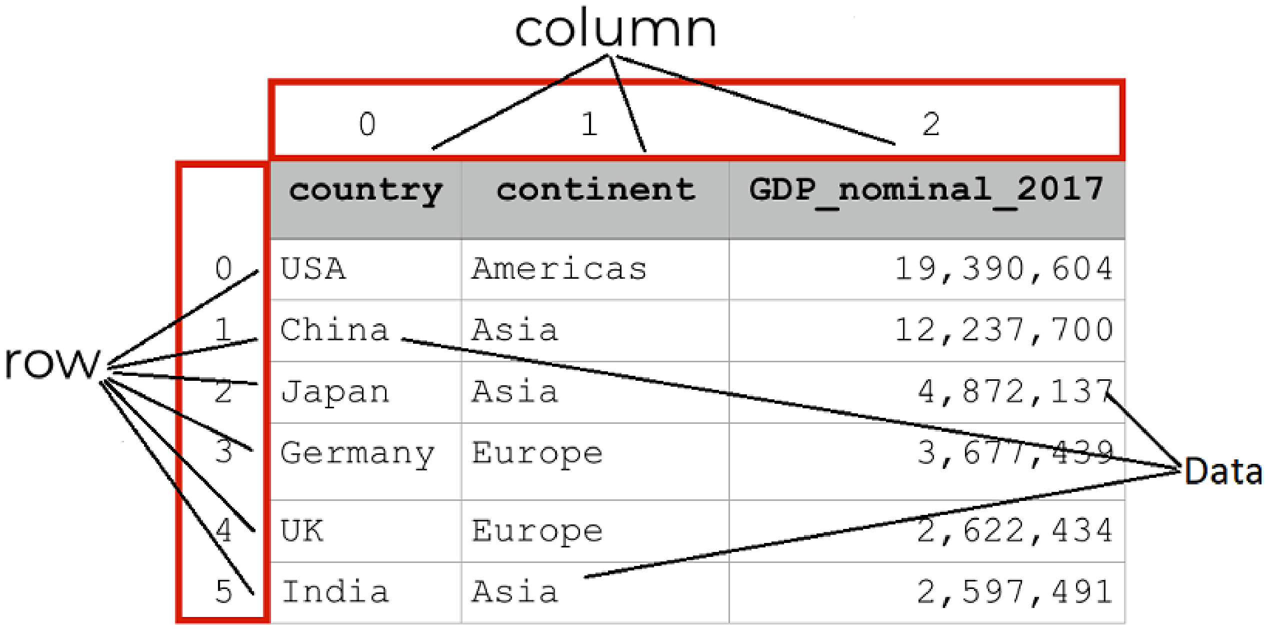
Pandas DataFrame

Pandas DataFrame is a two-dimensional data structure with labeled axes (rows and columns).

It consists of three principal components: the **data**, **rows** and **columns**.



	country	continent	GDP_nominal_2017
0	USA	Americas	19,390,604
1	China	Asia	12,237,700
2	Japan	Asia	4,872,137
3	Germany	Europe	3,677,439
4	UK	Europe	2,622,434
5	India	Asia	2,597,491

Topics:
– Create a Pandas DataFrame
– Rows and Columns Handling
– Subset the DataFrame

Create a Pandas DataFrame

In the real world, a Pandas DataFrame will be created by loading the datasets from existing storage. The storage can be CSV file, Excel file and SQL Database etc.

Pandas DataFrame can be created from the lists, dictionary, and from a list of dictionary etc.

Dataframe can be created in different ways and here, we have discussed some ways by which we create a dataframe:

Creating a dataframe from List

DataFrame can be created using a single list or a list of lists.

```
# import pandas as pd
import pandas as pd

# list of strings
lst = ['Code', 'Data', 'AI', 'is',
            'welcoming', 'all', 'Data Enthusiasts']

# Calling DataFrame constructor on list
df = pd.DataFrame(lst)
print(df)
```

Output

```
                  0
0              Code
1              Data
2                AI
3                is
4         welcoming
5               all
6  Data Enthusiasts
```

Creating DataFrame from dictionary of lists

Suppose you want to create a dataframe out of the below dictionary **dict**:

dict = {‘name’:[“Sunil”, “pankaj”, “sudhir”, “Geeku”],

‘degree’: [“MBA”, “BCA”, “M.Tech”, “MBA”],

‘score’:[90, 40, 80, 98]}

Here, the dictionary **keys** ‘name’ ,’degree’ & ‘score’ will become Columns and the **values (lists)** will become data.

Therefore all the values [“Sunil”, “pankaj”, “sudhir”, “Geeku”], [“MBA”, “BCA”, “M.Tech”, “MBA”] and [90, 40, 80, 98] need to be of **same length**.(here, it is 4).

Therefore, the basic condition to convert a dictionary of lists to a dataframe is that all the lists should be of same length.

```
# importing pandas as pd
import pandas as pd

# dictionary of lists
dict = {'name':["Sunil", "pankaj", "sudhir", "Geeku"],
        'degree': ["MBA", "BCA", "M.Tech", "MBA"],
        'score':[90, 40, 80, 98]}

df = pd.DataFrame(dict)

print(df)
```

Output

```
     name  degree  score
0   Sunil     MBA     90
1  pankaj     BCA     40
2  sudhir  M.Tech     80
3   Geeku     MBA     98
```

Rows and Columns Handling

A Data frame is a two-dimensional data structure where data are stored in rows and columns. Each row is called observation and each coulmn is termed as feature.

Here, you will perform basic operations on rows/columns like **selecting**, **adding**, **deleting** and **renaming**.

Select Rows

Pandas provide a unique method to retrieve rows from a Data frame. **`DataFrame.loc[]`** method is used to retrieve rows from Pandas DataFrame.

Rows can also be selected by passing integer location to an **iloc[]** function.

loc gets rows (**or** columns) with particular labels from the index.

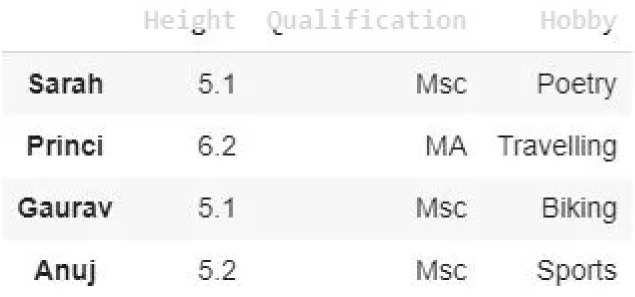
iloc gets rows (**or** columns) at particular positions in the index (it only takes integers)

```
# Import pandas package
import pandas as pd

# Define a dictionary containing Students data
data = {'Height': [5.1, 6.2, 5.1, 5.2],
        'Qualification': ['Msc', 'MA', 'Msc', 'Msc'],
        'Hobby': ['Poetry', 'Travelling', 'Biking', 'Sports']}

df = pd.DataFrame(data)
df.index = ['Sarah', 'Princi', 'Gaurav', 'Anuj']

print(df)

print ('\n Select Height, Qualification & Hobby for Sarah \n')
df.loc['Sarah']
```

Output

```
        Height Qualification       Hobby
Sarah      5.1           Msc      Poetry
Princi     6.2            MA  Travelling
Gaurav     5.1           Msc      Biking
Anuj       5.2           Msc      Sports

 Select Height, Qualification & Hobby for Sarah

Height               5.1
Qualification        Msc
Hobby             Poetry
Name: Sarah, dtype: object
```

Selecting rows by **index names**



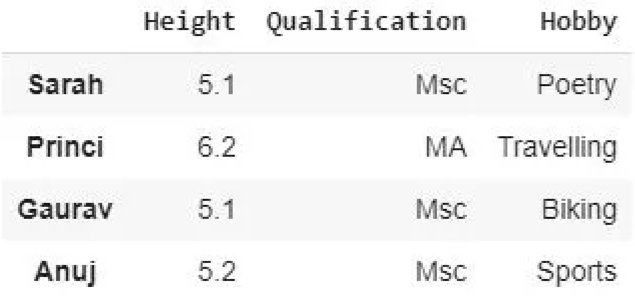
	Height	Qualification	Hobby
Sarah	5.1	Msc	Poetry
Princi	6.2	MA	Travelling
Gaurav	5.1	Msc	Biking
Anuj	5.2	Msc	Sports

```
# select two rows
df.loc[['Sarah','Gaurav']]
```

Output

```
    Height  Qualification   Hobby
Sarah   5.1 Msc      Poetry
Gaurav  5.1 Msc      Biking
```

Selecting rows by **slice of index names**



	Height	Qualification	Hobby
Sarah	5.1	Msc	Poetry
Princi	6.2	MA	Travelling
Gaurav	5.1	Msc	Biking
Anuj	5.2	Msc	Sports

```
# select 1st 3 rows
df.loc['Sarah':'Gaurav']
```

Output

```
    Height  Qualification   Hobby
Sarah   5.1 Msc Poetry
Princi  6.2 MA  Travelling
Gaurav  5.1 Msc Biking
```

Add Rows

To add a Row in Pandas DataFrame, you can **concat** the old dataframe with new one.

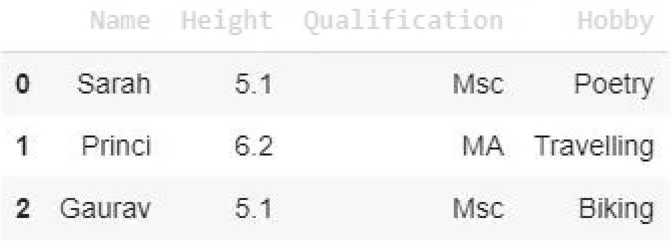
```
# Import pandas package
import pandas as pd

# Define a dictionary containing Students data
data = {'Height': [5.1, 6.2, 5.1, 5.2],
         'Qualification': ['Msc', 'MA', 'Msc', 'Msc'],
        'Hobby': ['Poetry', 'Travelling', 'Biking', 'Sports']}

df = pd.DataFrame(data)
df.index = ['Sarah', 'Princi', 'Gaurav', 'Anuj']

print(df)


new_row = pd.DataFrame({'Height': 5.3, 'Qualification': 'Bachelor' , 'Hobby': 'Sports'},
                                                            index =['NewRow'])
df = pd.concat([new_row, df])

print ('\n After adding row NewRow \n')

df
```

Output

```
       Height Qualification       Hobby
Sarah      5.1           Msc      Poetry
Princi     6.2            MA  Travelling
Gaurav     5.1           Msc      Biking
Anuj       5.2           Msc      Sports

 After adding row NewRow

       Height   Qualification   Hobby
NewRow  5.3 Bachelor    Sports
Sarah   5.1 Msc         Poetry
Princi  6.2 MA          Travelling
Gaurav  5.1 Msc         Biking
Anuj    5.2 Msc         Sports
```

You can also append all the rows of a dataframe to a new dataframe.

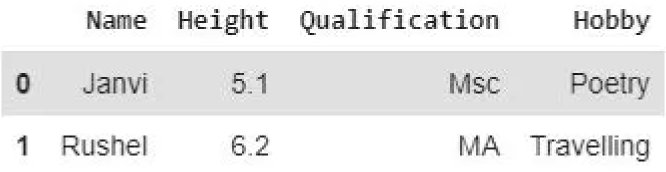
```
# Import pandas package
import pandas as pd

# Define a dictionary containing Students data
data = {'Name': ['Sarah', 'Princi', 'Gaurav'],
        'Height': [5.1, 6.2, 5.1],
        'Qualification': ['Msc', 'MA', 'Msc'],
        'Hobby': ['Poetry', 'Travelling', 'Biking']}

# Convert the dictionary into DataFrame
df = pd.DataFrame(data)
```



	Name	Height	Qualification	Hobby
0	Sarah	5.1	Msc	Poetry
1	Princi	6.2	MA	Travelling
2	Gaurav	5.1	Msc	Biking

Original DataFrame

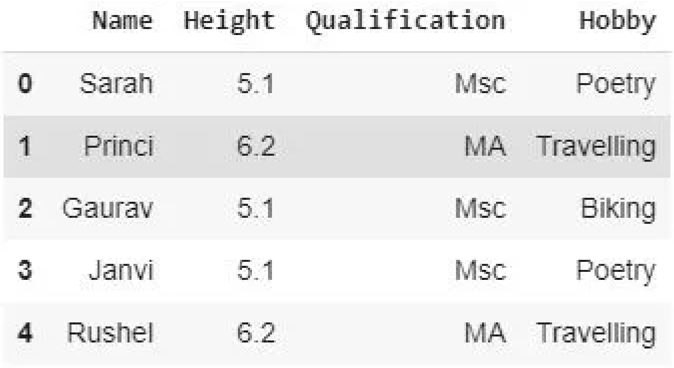
```
# Define a dictionary containing New data to append
data = {'Name': ['Janvi', 'Rushel'],
        'Height': [5.1, 6.2],
        'Qualification': ['Msc', 'MA'],
        'Hobby': ['Poetry', 'Travelling']}

# Convert the dictionary into DataFrame
df_to_append = pd.DataFrame(data)
```



	Name	Height	Qualification	Hobby
0	Janvi	5.1	Msc	Poetry
1	Rushel	6.2	MA	Travelling

DataFrame to Append

```
#append the df_to_append to the original dataframe
df.append(df_to_append, ignore_index= True)
```



	Name	Height	Qualification	Hobby
0	Sarah	5.1	Msc	Poetry
1	Princi	6.2	MA	Travelling
2	Gaurav	5.1	Msc	Biking
3	Janvi	5.1	Msc	Poetry
4	Rushel	6.2	MA	Travelling

DataFrame after Append

Delete Rows

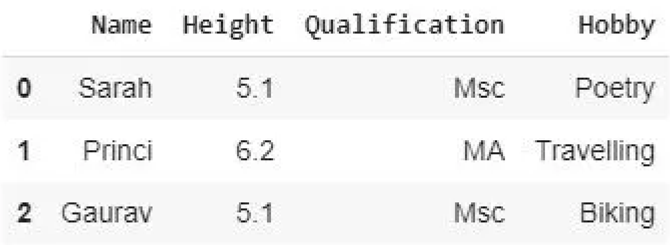
To delete a row in Pandas DataFrame, we can use the drop() method.

Rows is deleted by dropping Rows by index label.

```
# Import pandas package
import pandas as pd

# Define a dictionary containing Students data
```

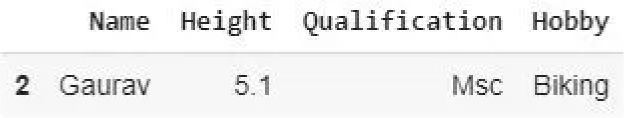
```
data = {'Name': ['Sarah', 'Princi', 'Gaurav'],
        'Height': [5.1, 6.2, 5.1],
        'Qualification': ['Msc', 'MA', 'Msc'],
        'Hobby': ['Poetry', 'Travelling', 'Biking']}

# Convert the dictionary into DataFrame
df = pd.DataFrame(data)

print(df)

print('\n After dropping \n')

df.drop([0,1])
```

Output



	Name	Height	Qualification	Hobby
0	Sarah	5.1	Msc	Poetry
1	Princi	6.2	MA	Travelling
2	Gaurav	5.1	Msc	Biking

Before Dropping Rows



	Name	Height	Qualification	Hobby
2	Gaurav	5.1	Msc	Biking

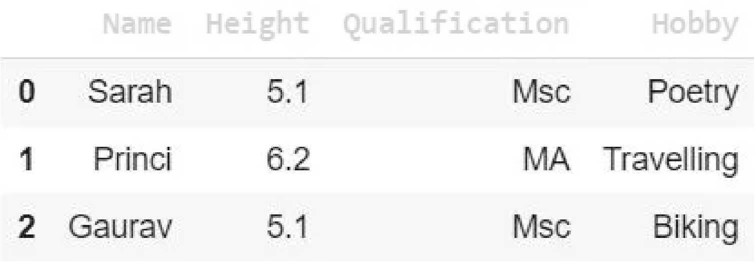
After dropping 1st 2 rows

Rename Index

You can rename index names **rename()** function.

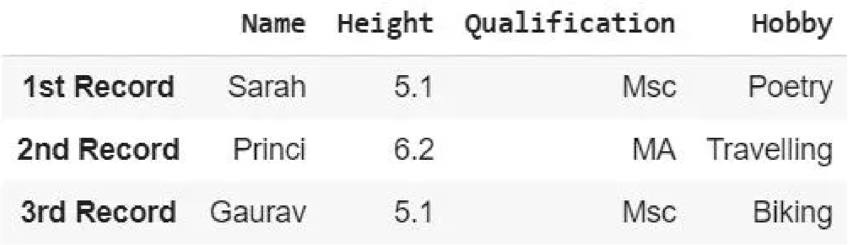
```
import pandas as pd

# making data frame from csv file
# Define a dictionary containing Students data
data = {'Name': ['Sarah', 'Princi', 'Gaurav'],
        'Height': [5.1, 6.2, 5.1],
        'Qualification': ['Msc', 'MA', 'Msc'],
        'Hobby': ['Poetry', 'Travelling', 'Biking']}

# Convert the dictionary into DataFrame
df = pd.DataFrame(data)
```

```
# changing index cols with rename() to 1st, 2nd & 3rd Record respectively
df.rename(index = {  0: "1st Record",
                     1:"2nd Record",
                     2: "3rd Record"},
                                 inplace = True)
# display
df
```



	Name	Height	Qualification	Hobby
0	Sarah	5.1	Msc	Poetry
1	Princi	6.2	MA	Travelling
2	Gaurav	5.1	Msc	Biking

Before Renaming Index



	Name	Height	Qualification	Hobby
1st Record	Sarah	5.1	Msc	Poetry
2nd Record	Princi	6.2	MA	Travelling
3rd Record	Gaurav	5.1	Msc	Biking

After Renaming Index

Select Column

To select a column in Pandas DataFrame, you can either access the columns by calling them by their columns name or column number.

Select columns by **column name**

```
# Import pandas package
import pandas as pd

# Define a dictionary containing employee data
data = {'Name':['Sarah', 'Princi', 'Gaurav', 'Anuj'],
        'Age':[27, 24, 22, 32],
        'Address':['Kolkata', 'Kanpur', 'Allahabad', 'Delhi'],
        'Qualification':['Msc', 'MA', 'MCA', 'Phd']}

# Convert the dictionary into DataFrame
df = pd.DataFrame(data)

# select two columns
df[['Name', 'Qualification']]
```

Output

```
    Name    Qualification
0   Sarah   Msc
1   Princi  MA
2   Gaurav  MCA
3   Anuj    Phd
```

Select columns by **column number**

```
# Import pandas package
import pandas as pd

# Define a dictionary containing employee data
data = {'Name':['Sarah', 'Princi', 'Gaurav', 'Anuj'],
        'Age':[27, 24, 22, 32],
        'Address':['Kolkata', 'Kanpur', 'Allahabad', 'Delhi'],
        'Qualification':['Msc', 'MA', 'MCA', 'Phd']}

# Convert the dictionary into DataFrame
df = pd.DataFrame(data)


# select all rows by ':'
# select two columns Name and Qualification by their position

df.iloc[:, [0,3]]
```

Output

```
    Name    Qualification
0   Sarah   Msc
1   Princi  MA
2   Gaurav  MCA
3   Anuj    Phd
```

Select **1st 3 columns**

```
df[df.columns[0:3]]
```

Output

```
    Name    Age Address
0   Sarah   27  Kolkata
1   Princi  24  Kanpur
2   Gaurav  22  Allahabad
3   Anuj    32  Delhi
```

Select columns from “**Name**” to “**Address**”

```
# select two rows and
# column "name" to "Address"
# Means total three columns
df.loc[ :,'Name':'Address']
```

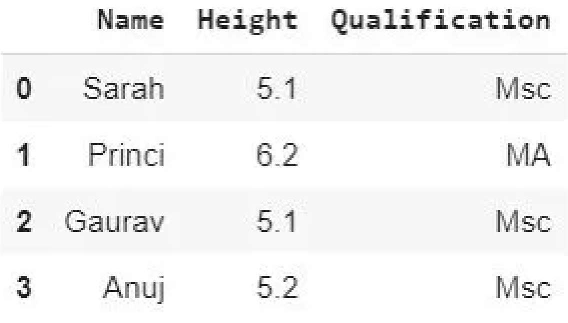
Output

```
    Name    Age Address
0   Sarah   27  Kolkata
1   Princi  24  Kanpur
2   Gaurav  22  Allahabad
3   Anuj    32  Delhi
```

Add Column

To add a column in Pandas DataFrame, you can declare a new list as a column and add to a existing Dataframe.

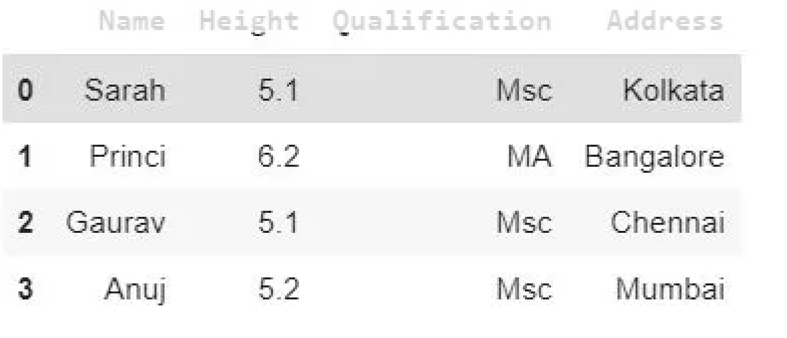
```
# Import pandas package
import pandas as pd

# Define a dictionary containing Students data
data = {'Name': ['Sarah', 'Princi', 'Gaurav', 'Anuj'],
        'Height': [5.1, 6.2, 5.1, 5.2],
        'Qualification': ['Msc', 'MA', 'Msc', 'Msc']}

# Convert the dictionary into DataFrame
df = pd.DataFrame(data)
df
```



	Name	Height	Qualification
0	Sarah	5.1	Msc
1	Princi	6.2	MA
2	Gaurav	5.1	Msc
3	Anuj	5.2	Msc

Before Adding Column

```
# Declare a list that is to be converted into a column
address = ['Kolkata', 'Bangalore', 'Chennai', 'Mumbai']

# Using 'Address' as the column name
# and equating it to the list
df['Address'] = address
df
```



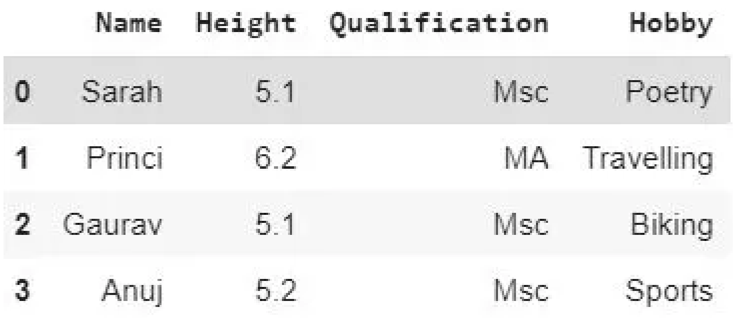
	Name	Height	Qualification	Address
0	Sarah	5.1	Msc	Kolkata
1	Princi	6.2	MA	Bangalore
2	Gaurav	5.1	Msc	Chennai
3	Anuj	5.2	Msc	Mumbai

After Adding Column Address

Delete Column

To delete a column in Pandas DataFrame, you can use the `drop()` method. Columns are deleted by **dropping** columns with columnnames.

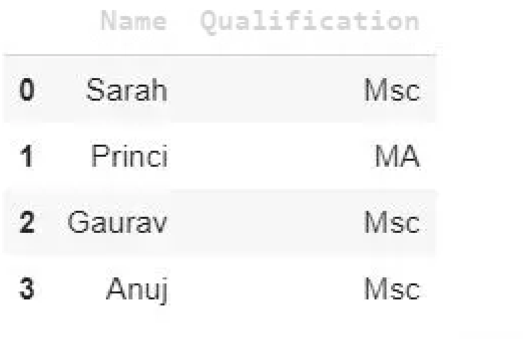
```
# Import pandas package
import pandas as pd

# Define a dictionary containing Students data
data = {'Name': ['Sarah', 'Princi', 'Gaurav', 'Anuj'],
        'Height': [5.1, 6.2, 5.1, 5.2],
        'Qualification': ['Msc', 'MA', 'Msc', 'Msc'],
        'Hobby': ['Poetry', 'Travelling', 'Biking', 'Sports']}

# Convert the dictionary into DataFrame
df = pd.DataFrame(data)
df
```



	Name	Height	Qualification	Hobby
0	Sarah	5.1	Msc	Poetry
1	Princi	6.2	MA	Travelling
2	Gaurav	5.1	Msc	Biking
3	Anuj	5.2	Msc	Sports

Before Deleting Columns

```
df.drop(["Height", "Hobby"], axis = 1, inplace = True)
```



	Name	Qualification
0	Sarah	Msc
1	Princi	MA
2	Gaurav	Msc
3	Anuj	Msc

After Deleting Columns

Rename Column

To Rename columns, you can use **df.columns** = **new list of column names**.

```
# Import pandas package
import pandas as pd

# Define a dictionary containing Students data
data = {'Name': ['Sarah', 'Princi', 'Gaurav', 'Anuj'],
        'Height': [5.1, 6.2, 5.1, 5.2],
        'Qualification': ['Msc', 'MA', 'Msc', 'Msc'],
        'Hobby': ['Poetry', 'Travelling', 'Biking', 'Sports']}

# Convert the dictionary into DataFrame
df = pd.DataFrame(data)
df
# change the column name Qualification & Hobby to Degree and Leisure respectively
df.columns = ['Name','Height', 'Degree', 'Leisure']
df
```

Output

```
        Name    Height  Degree  Leisure
0   Sarah   5.1 Msc Poetry
1   Princi  6.2 MA  Travelling
2   Gaurav  5.1 Msc Biking
3   Anuj    5.2 Msc Sports
```

You can choose to **rename a particular column** as well.

```
# renaming 'Degree' back to 'Qualification'
df.rename(columns={'Degree':'Qualification'}, inplace=True)
df
```

```
    Name    Height  Qualification   Leisure
0   Sarah   5.1 Msc         Poetry
```

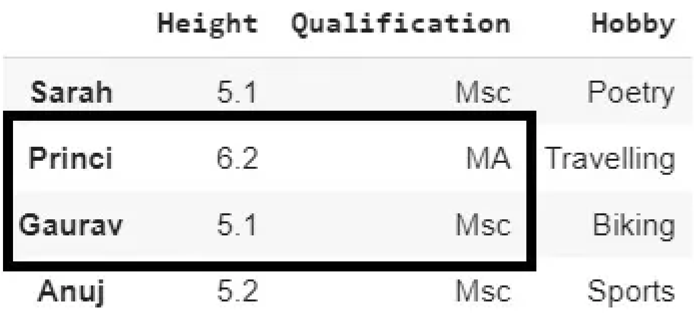
```
1   Princi   6.2 MA         Travelling
2   Gaurav   5.1 Msc        Biking
3   Anuj     5.2 Msc        Sports
```

Subset the DataFrame

Subsetting a Dataframe is same as slicing a dataframe into smaller dataframe so that you can focus on a small chunk of a large dataset at a particular time.

It is one of the main skills for ***Exploratory Data Analysis*** (EDA).

For example, you need to slice the box out of the dataframe:

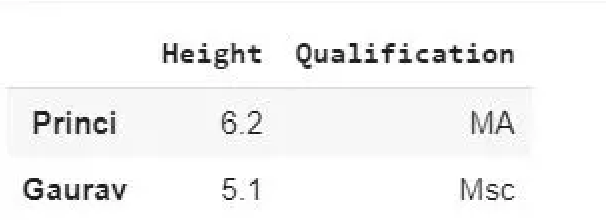


	Height	Qualification	Hobby
Sarah	5.1	Msc	Poetry
Princi	6.2	MA	Travelling
Gaurav	5.1	Msc	Biking
Anuj	5.2	Msc	Sports

Subsetting the Squared Portion

To do this, you will subset the two rows (**2nd** and **3rd**) and 1st two coulmns (**Height** & **Qualification**).

```
df.iloc[1:3,0:2]
```



	Height	Qualification
Princi	6.2	MA
Gaurav	5.1	Msc