

Python

Case Study - Pandas

Table of Contents

1. **Pre-requisites**
2. **Problem Statement**
3. **Dataset**
4. **Opening IDE**
5. **Creating Python File**
6. **Importing Pandas Module**
7. **Solving the Queries**

**Pre-requisites:**

1. Python Installed.
2. Excel.

**Problem Statement:**

1. Explain about Pandas and why do we use them?
2. What can Pandas do?
3. And do the following with the given CSV file

1.Data Type conversions

-- int to float.

-- string to int.

2.Reindexing columns.

3.Read data from nth row from file

4.Read data from different sheets in an Excel file.

5.Remove columns

6.Add new columns

7.Filter data

-- based on a single/ multiple values

-- based on a number range

8.Split data from single column to multiple columns using a delimiter

9.Join data between two data frames

-- explore different types of joins

1o.Rename column names in dataframe.

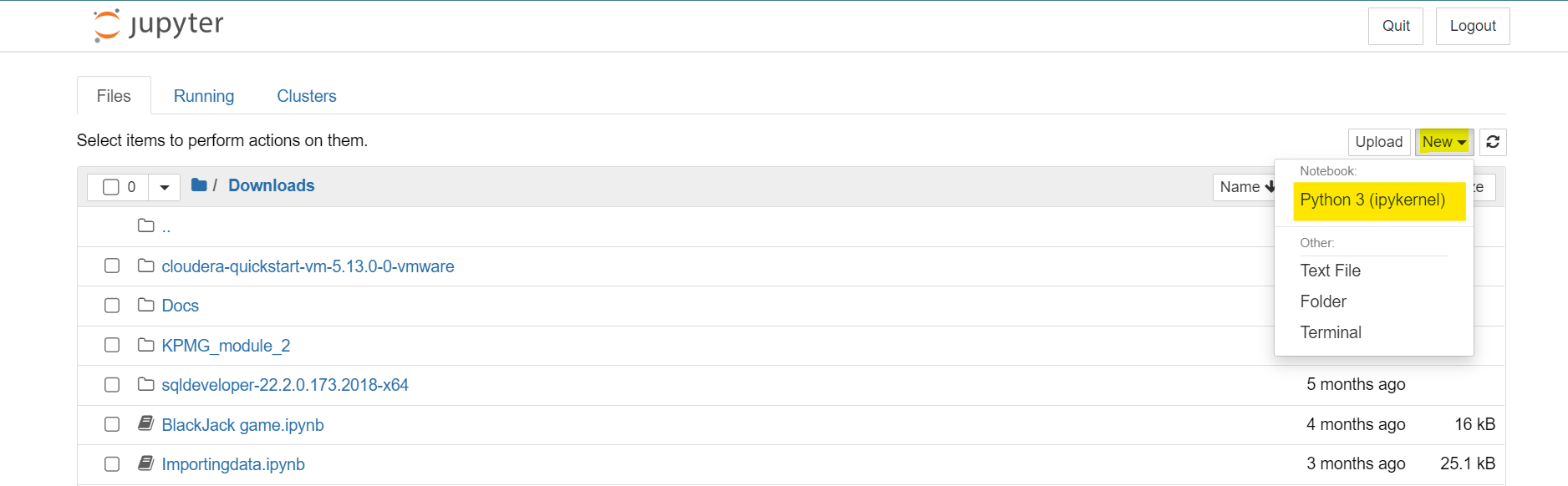
**Dataset:**

**The Dataset used for this project.**

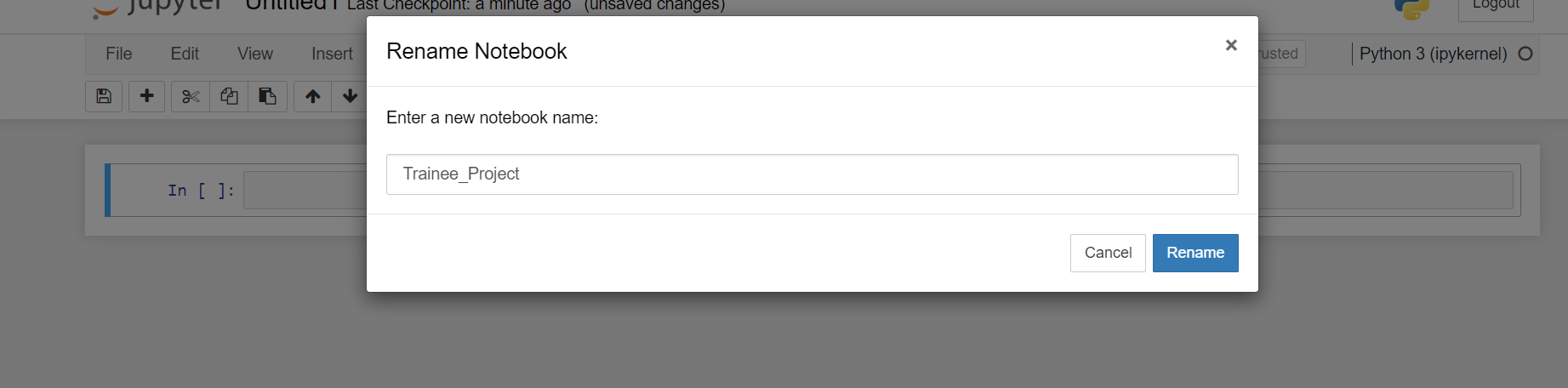


# Step 1 : Open Jupiter Notebook to work on.

Step 2 : Create a new python file.



Step 3 : Rename the python file.



## **What is Pandas?**

Pandas is a Python library used for working with data sets. It has functions for analysing, cleaning, exploring, and manipulating data.

The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

## **Why Use Pandas?**

Pandas allows us to analyse big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science.

## **What Can Pandas Do?**

Pandas gives you answers about the data. Like:

* Is there a correlation between two or more columns?
* What is average value?
* Max value?
* Min value?

Pandas are also able to delete rows that are not relevant, or contains wrong values, like empty or NULL values. This is called *cleaning* the data.

**Step 1:**

Importing pandas to use its functions.

**Code:** import pandas as pd



**Step 2:**

Importing the required dataset and storing it inside a dataframe ‘a’ and copy the data to dataframe ‘df’.

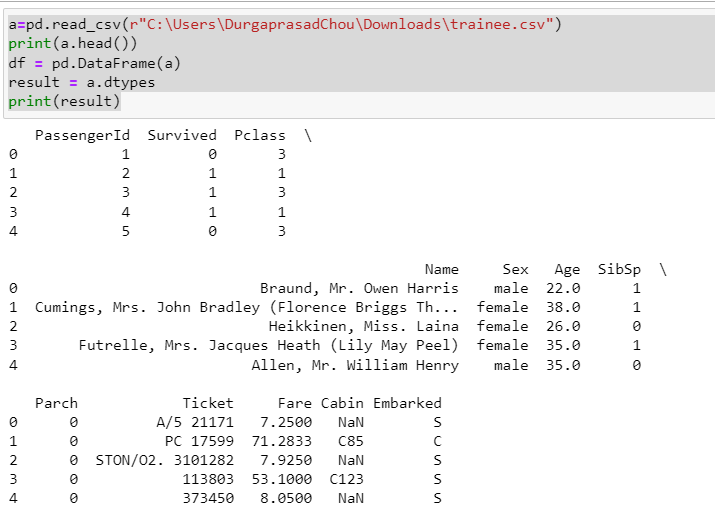
**Code:**a=pd.read\_csv(r"C:\Users\DurgaprasadChou\Downloads\trainee.csv")

print(a.head())

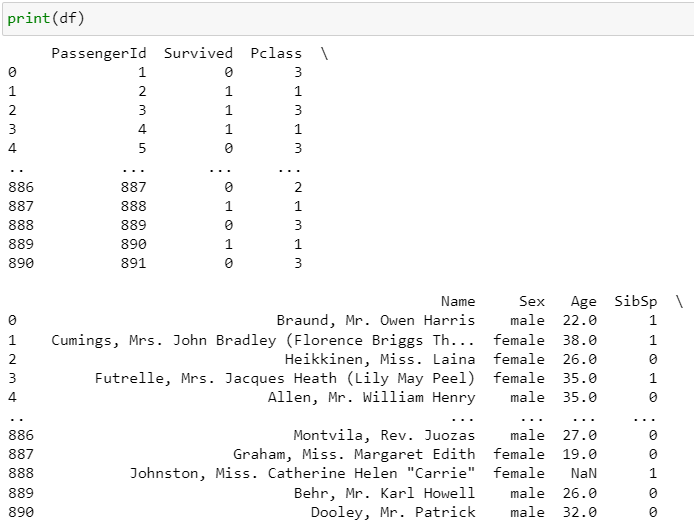
df = pd.DataFrame(a)

result = a.dtypes

print(result)



**Output:**

****

**Step 4:**

Changing the column datatype to the asked datatype.

1)#Changing Datatype from int to float.

2)#Changing Data type from string to int.

**#Changing Datatype from int to float.**

**Code:**  print("datatype before getting changed")

print(df["Pclass"])

print()

print()

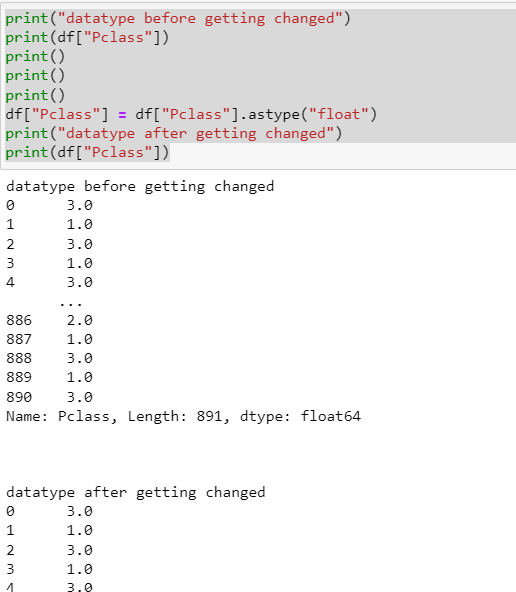
print()

df["Pclass"] = df["Pclass"].astype("float")

print("datatype after getting changed")

print(df["Pclass"])

**Output**



**#Changing Data Type from string to int.**

**Code:** print("datatype before getting changed")

print(df["Sex"])

print()

print()

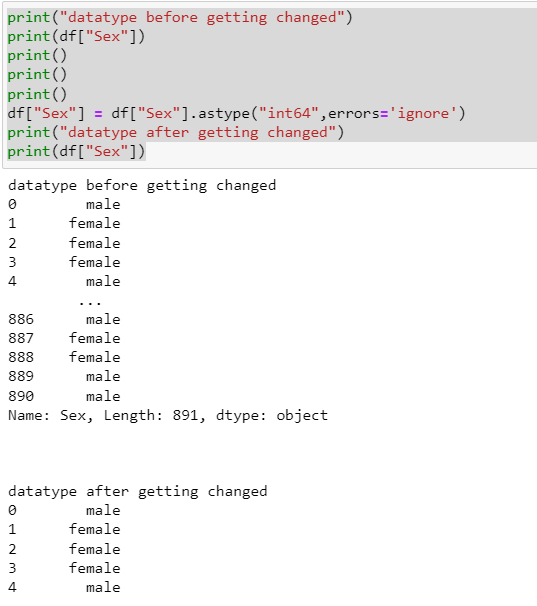
print()

df["Sex"] = df["Sex"].astype("int64",errors='ignore')

print("datatype after getting changed")

print(df["Sex"])

**Output:**



**Step 5:**

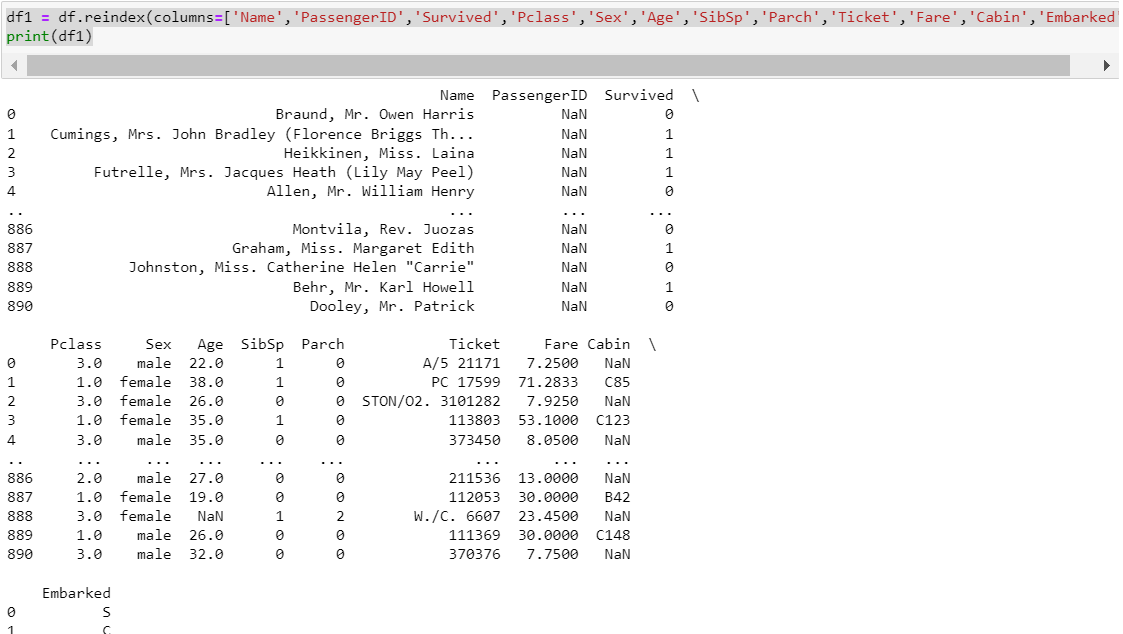
Reindexing column.

Bringing the name column in the front.

**Code:** df1 = df.reindex(columns=['Name','PassengerID','Survived','Pclass','Sex','Age','SibSp','Parch','Ticket','Fare','Cabin','Embarked'])

print(df1)

**Output:**

****

**Step 6:**

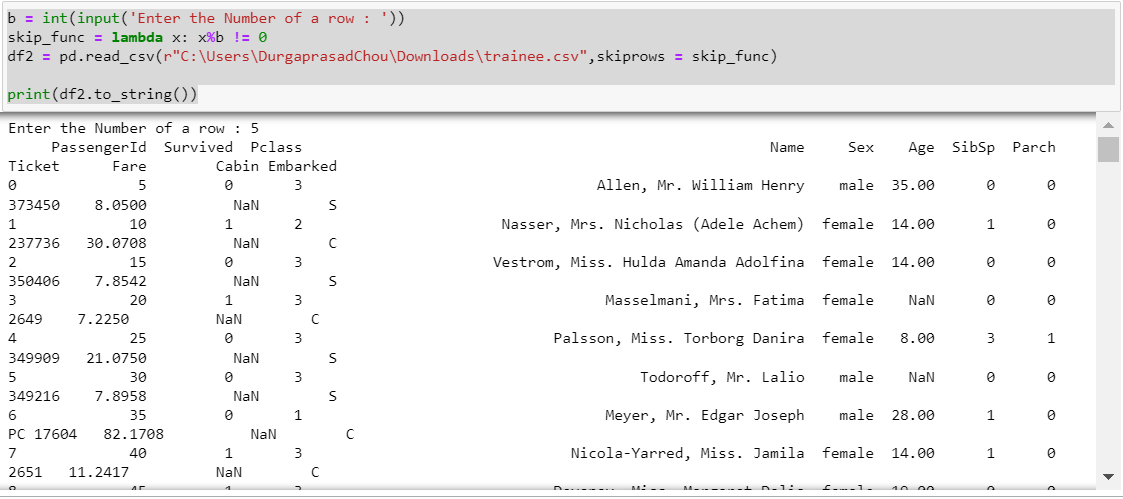
Read data from the nth row of the column

**Code:**  b = int(input('Enter the Number of a row : '))

skip\_func = lambda x: x%b != 0

df2 = pd.read\_csv(r"C:\Users\DurgaprasadChou\Downloads\trainee.csv",skiprows = skip\_func)

print(df2.to\_string())

****

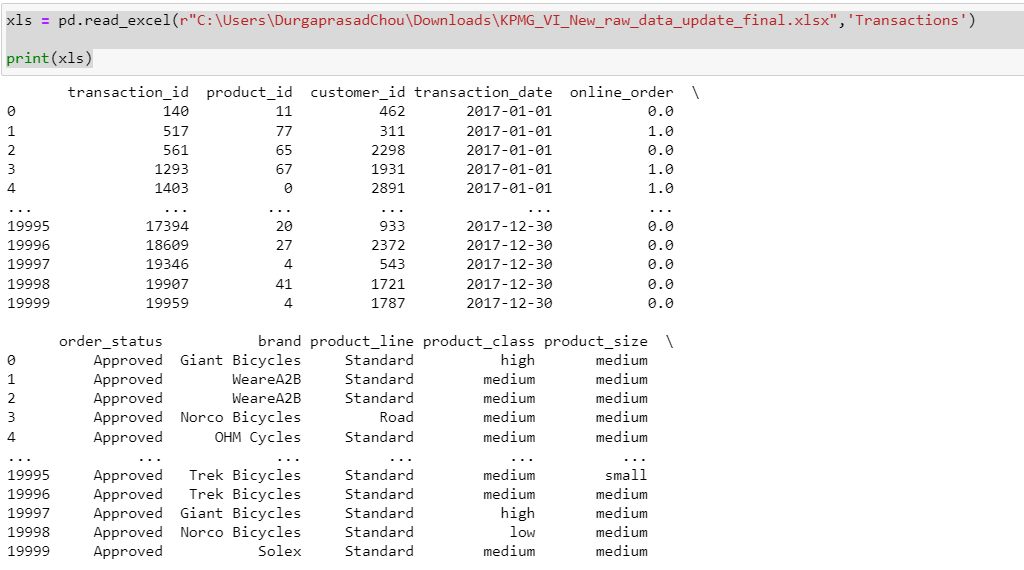
**Step 7:**

Reading data from a different sheet from a excel file.

**Code:** xls = pd.read\_excel(r"C:\Users\DurgaprasadChou\Downloads\KPMG\_VI\_New\_raw\_data\_update\_final.xlsx",'Transactions')

print(xls)

**Output:**



**Step 8:**

Removing a column.

1)#Deleting a singular column.

2)#Deleting multiple columns.

**#Deleting a singular column.**

**Code:**  #Remove columns

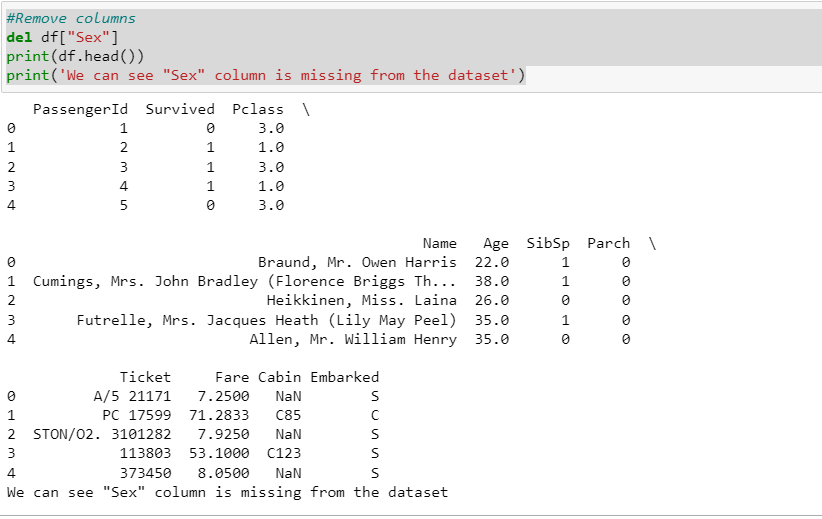
del df["Sex"]

print(df.head())

print('We can see "Sex" column is missing from the dataset')

**Output:**

**Output:**



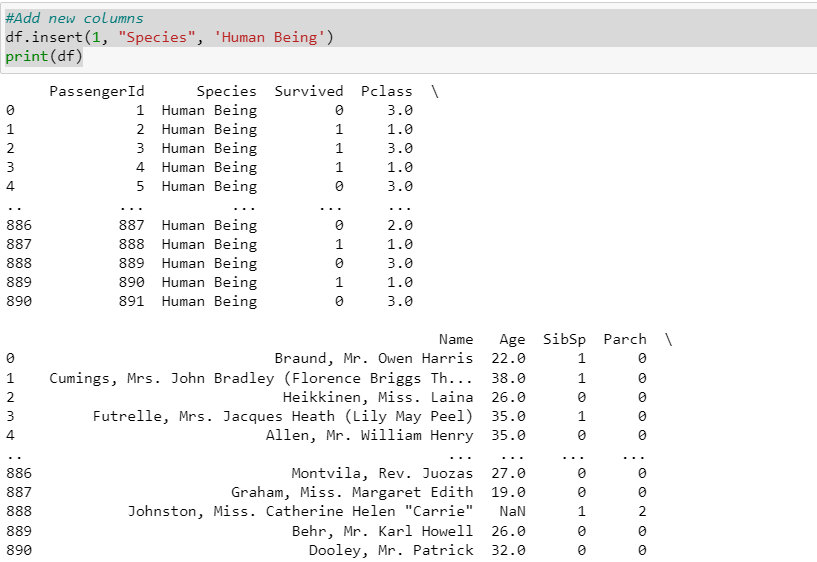
#Adding a column

**Code:** #Add new columns

df.insert(1, "Species", 'Human Being')

print(df)

**Output:**

****

**Step 10:**

Filtering Data based on:

1)#Singular column.

2)#Range based.

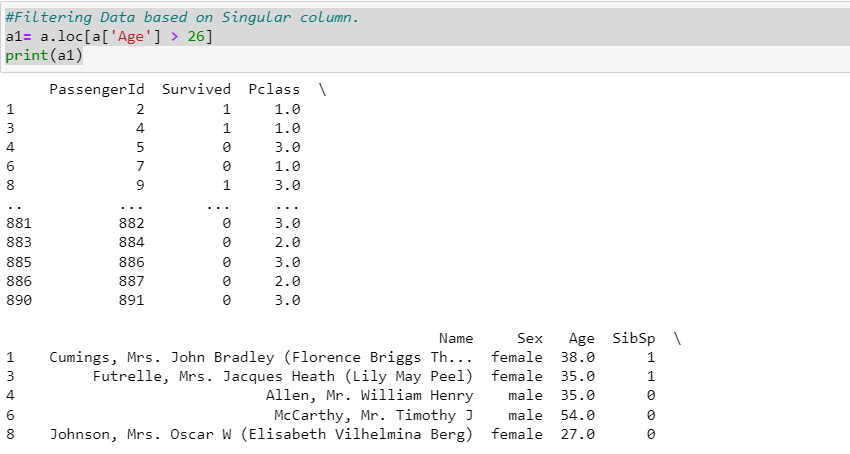
**#Singular column.**

**Code:** #Filtering Data based on Singular column.

a1= a.loc[a['Age'] > 26]

print(a1)

**Output:**

****

**#Range based.**

**Code:** #Filtering Data based on based on a number range.

a.loc[(a['Age'] >= 25) & (a['Age'] <= 50)]



**Step 11:**

Split data from single column to multiple column.

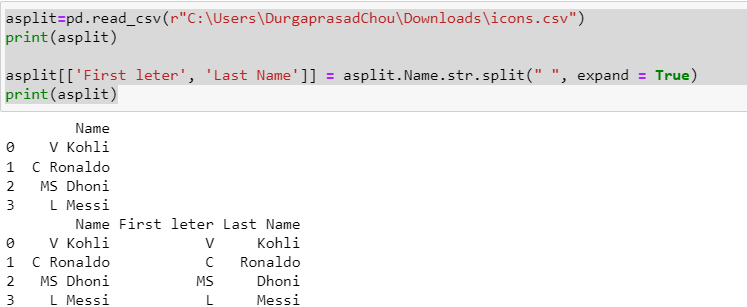
**Code:** asplit=pd.read\_csv(r"C:\Users\DurgaprasadChou\Downloads\icons.csv")

print(asplit)

asplit[['First leter', 'Last Name']] = asplit.Name.str.split(" ", expand = True)

print(asplit)

**Output:**



**Step 12:**

Join data between two data frames

**Code:** i1 = pd.read\_csv(r"C:\Users\DurgaprasadChou\Downloads\Icon1.csv")

print(i1.head())

print()

print()

i2 = pd.read\_csv(r"C:\Users\DurgaprasadChou\Downloads\Icon2.csv")

print(i2)

print('Lets perform Concatination')

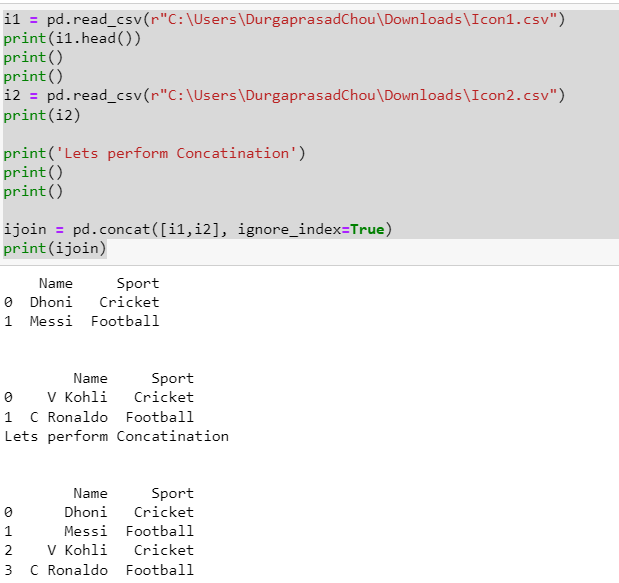
print()

print()

ijoin = pd.concat([i1,i2], ignore\_index=True)

print(ijoin)

**Output:**

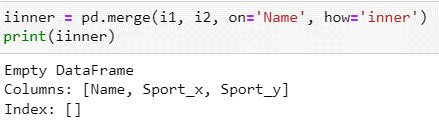
****

**#Inner Join**

**Code:** iinner = pd.merge(i1, i2, on='Name', how='inner')

print(iinner)

**Output:**

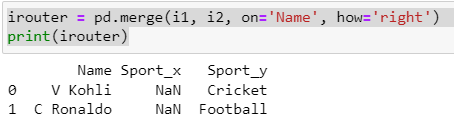


**#Right Outer Join**

**Code:** irouter = pd.merge(i1, i2, on='Name', how='right')

print(irouter)

**Output:**

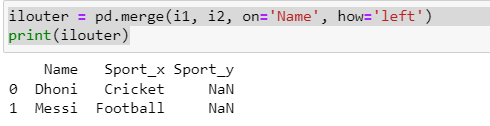
****

**#Left Outer Join**

**Code:** ilouter = pd.merge(i1, i2, on='Name', how='left')

print(ilouter)

**Output:**

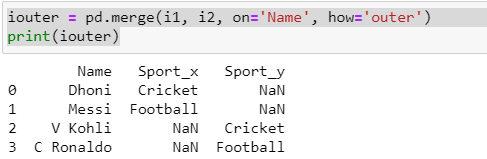
****

**#Outer Join**

**Code:** iouter = pd.merge(i1, i2, on='Name', how='outer')

print(iouter)

**Output:**

****

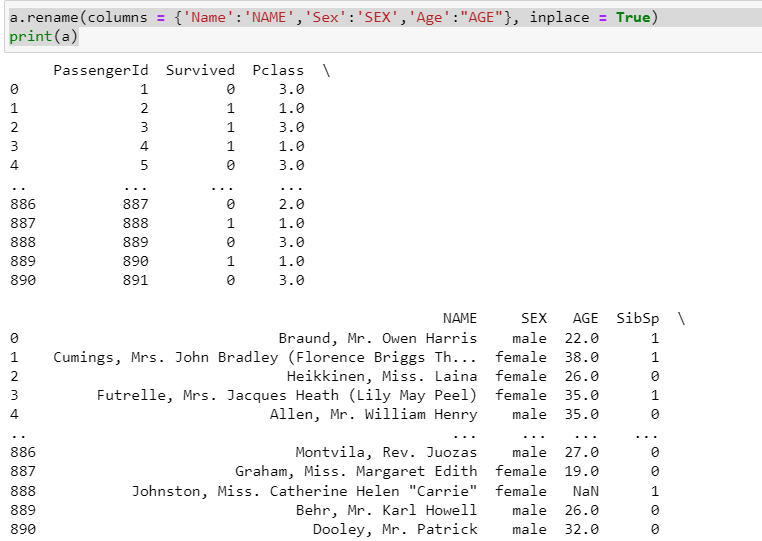
**Step 13:**

Renaming columns name in dataframe.

**Code:** a.rename(columns = {'Name':'NAME','Sex':'SEX','Age':"AGE"}, inplace = True)

print(a)

**Output:**

****

****

**Attaching case study notebook for Reference**