Experiment 10: Reading Spreadsheets

Aim: To read CSV and Excel files using Python.

Theory:

Introduction to Pandas:

Pandas is a powerful python data analysis toolkit for reading, filtering, manipulating, visualizing and exporting data. It has a library that is required for processing data very efficiently.

Pandas provide wide range of functionalities such as:

- It can read variety of data. Eg: csv, excel, json, etc.
- It has functions for filtering, selecting and manipulating data.
- It plots data for visualization and exploration purpose.
- It has huge contribution from the developer community.

Reading a spreadsheet using Pandas:

Pandas can read wide varieties of files such as:

text	CSV, JSON, HTML, local clipboard
binary	MS Excel, HDF5 Format, Feather Format, Msgpack, Stata, SAS, Python Pickle Format
SQL	SQL, Google Big Query

Basic Operations:

.read_csv(“filename.csv”) : Reads the file from the Folder.

.shape: Gives the count of rows and columns in the file.

.head(n): Gives the values mentioned in top ‘n’ rows. If ‘n’ is not mentioned, it assumes n=5.

.tail(n): Gives the values mentioned in last ‘n’ rows. If ‘n’ is not mentioned, it assumes n=5.

.columns: Gives the names of the columns.

.duplicated(): Gives Boolean output by comparing the data in the entire row.

.column_name.duplicated(): Gives Boolean output by comparing the data on in that column.

.duplicated().sum(): Adds the True values and gives the count of the duplicated data.

.loc[df.duplicated(), :]: Mentions the top rows that are being duplicated by default.

.loc[df.duplicated(keep='first'), :]: Mentions the rows that are being duplicated, it starts checking from top.

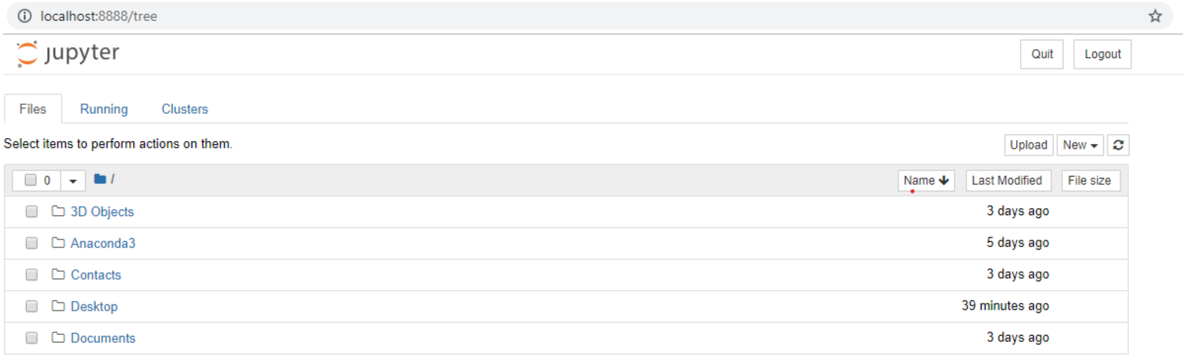
.loc[df.duplicated(keep='last'), :]: Mentions the rows that are being duplicated, it starts checking from bottom.

.drop_duplicates(): It eliminates the copied rows.

.drop_duplicates('column_name'): It eliminates the data that was copied in the mentioned column.

Steps:

1. Open Jupyter Notebook.

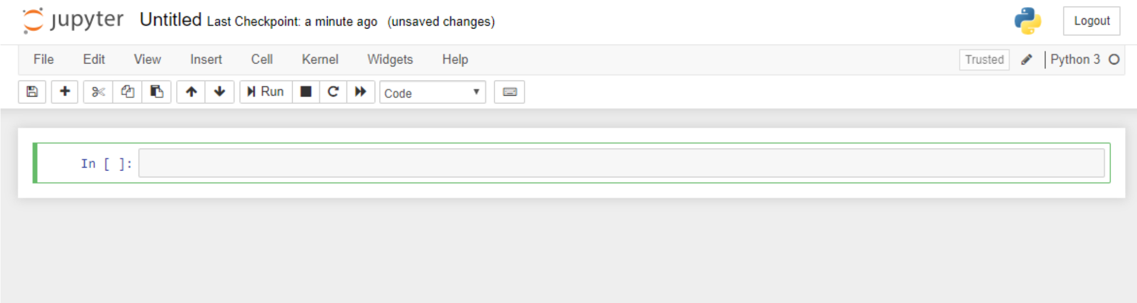


localhost:8888/tree

Select items to perform actions on them.

Name	Last Modified	File size
3D Objects	3 days ago	
Anaconda3	5 days ago	
Contacts	3 days ago	
Desktop	39 minutes ago	
Documents	3 days ago	

2. Open a New File by clicking New → Python3; a new python3 file opens, where we will be writing the codes.



Untitled Last Checkpoint: a minute ago (unsaved changes)

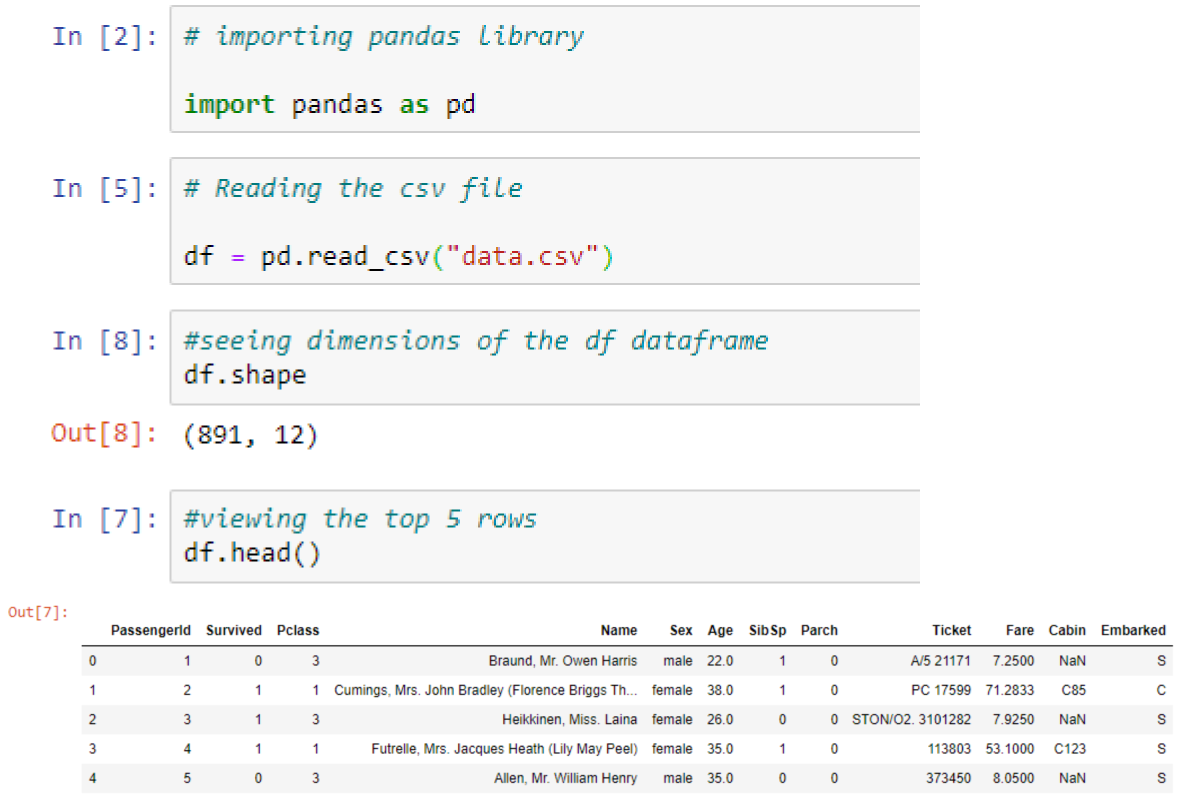
In [ ]:

3. Example Code: #importing required libraries *# indicates this is a comment*

   import pandas as pd *#imports pandas*

4. To obtain the result, press "**Ctrl+Enter**".

Code:



In [2]:
```
# importing pandas library

import pandas as pd
```

In [5]:
```
# Reading the csv file

df = pd.read_csv("data.csv")
```

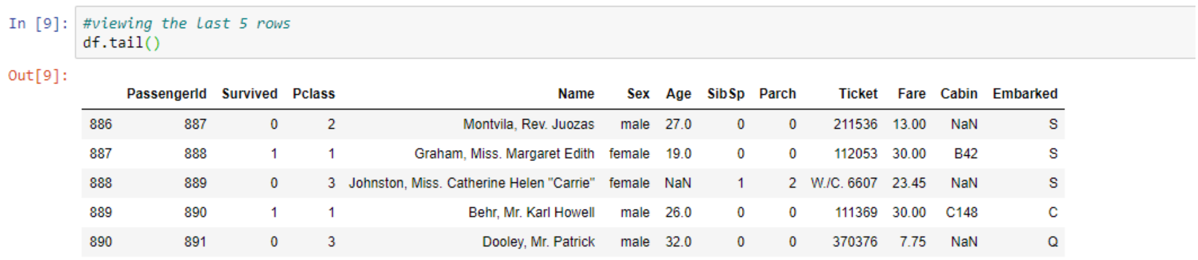
In [8]:
```
#seeing dimensions of the df dataframe
df.shape
```

Out[8]:
```
(891, 12)
```

In [7]:
```
#viewing the top 5 rows
df.head()
```

Out[7]:

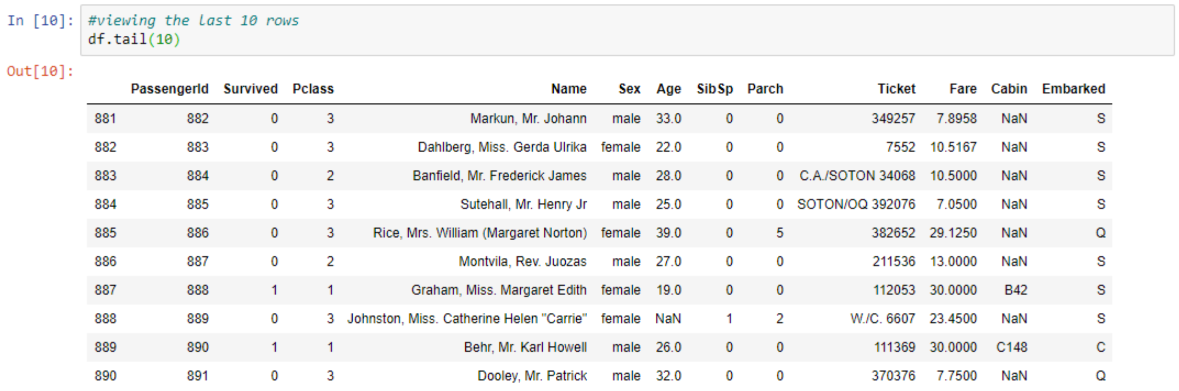
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S



In [9]:
```
#viewing the last 5 rows
df.tail()
```

Out[9]:

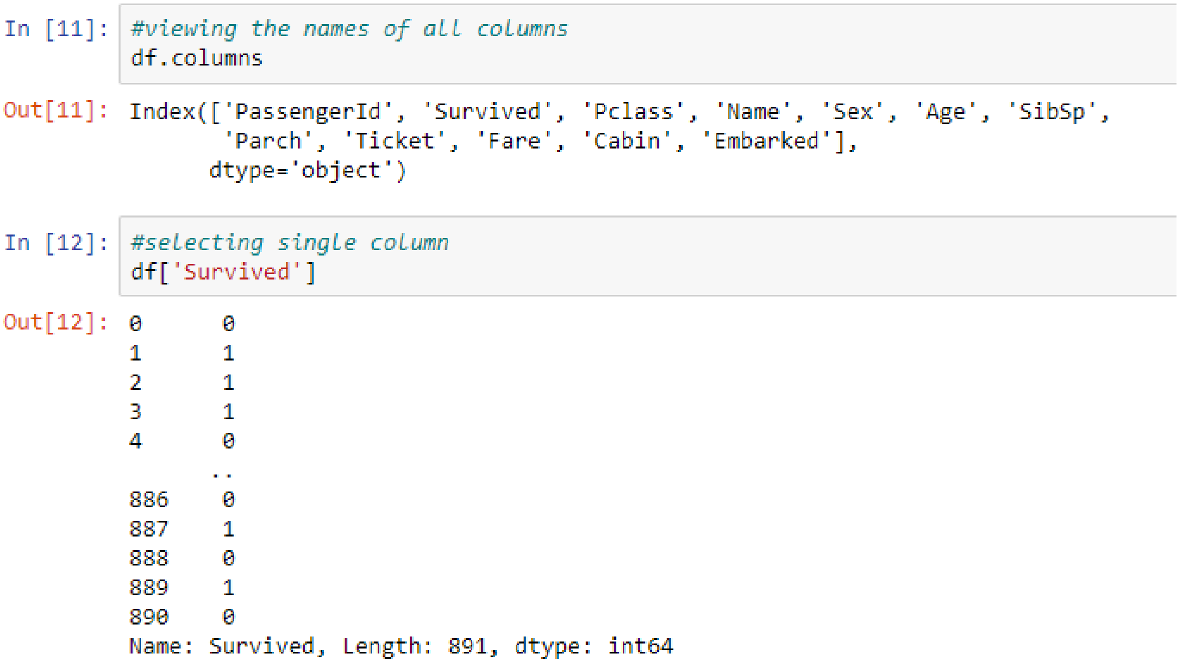
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q



In [10]:
```
#viewing the last 10 rows
df.tail(10)
```

Out[10]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
881	882	0	3	Markun, Mr. Johann	male	33.0	0	0	349257	7.8958	NaN	S
882	883	0	3	Dahlberg, Miss. Gerda Ulrika	female	22.0	0	0	7552	10.5167	NaN	S
883	884	0	2	Banfield, Mr. Frederick James	male	28.0	0	0	C.A./SOTON 34068	10.5000	NaN	S
884	885	0	3	Sutehall, Mr. Henry Jr	male	25.0	0	0	SOTON/OQ 392076	7.0500	NaN	S
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.1250	NaN	Q
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

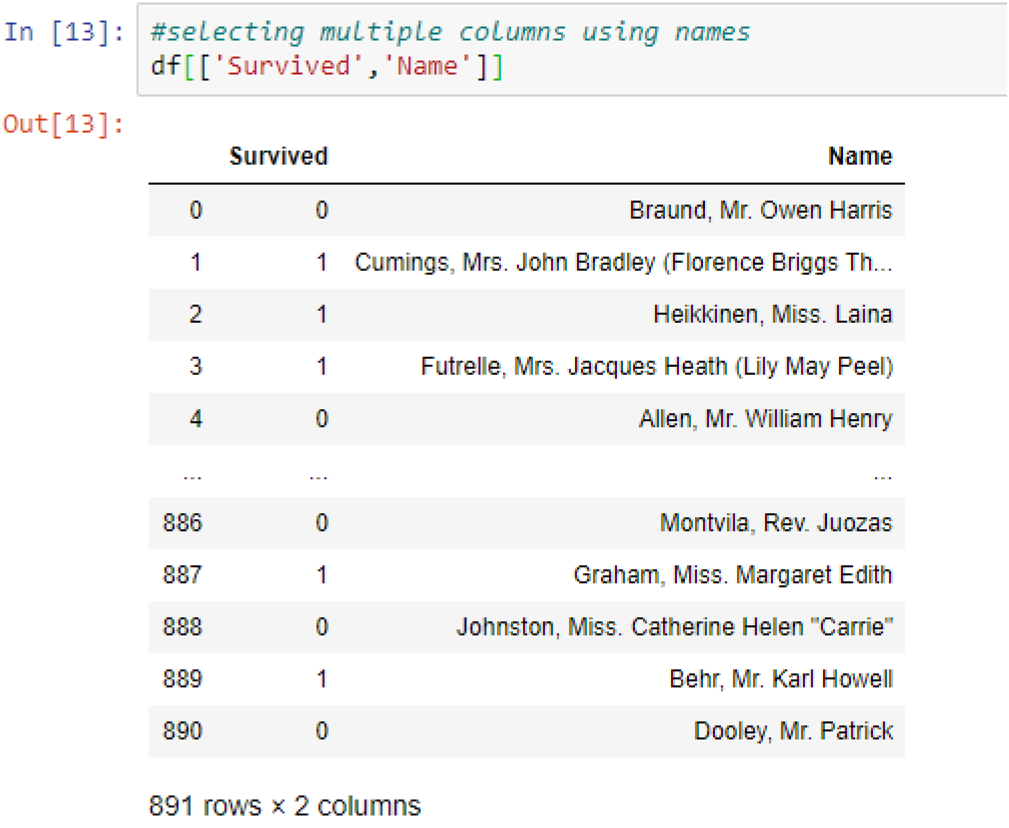


```
In [11]: #viewing the names of all columns
         df.columns

Out[11]: Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
                'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
               dtype='object')

In [12]: #selecting single column
         df['Survived']

Out[12]: 0      0
         1      1
         2      1
         3      1
         4      0
               ..
         886    0
         887    1
         888    0
         889    1
         890    0
         Name: Survived, Length: 891, dtype: int64
```

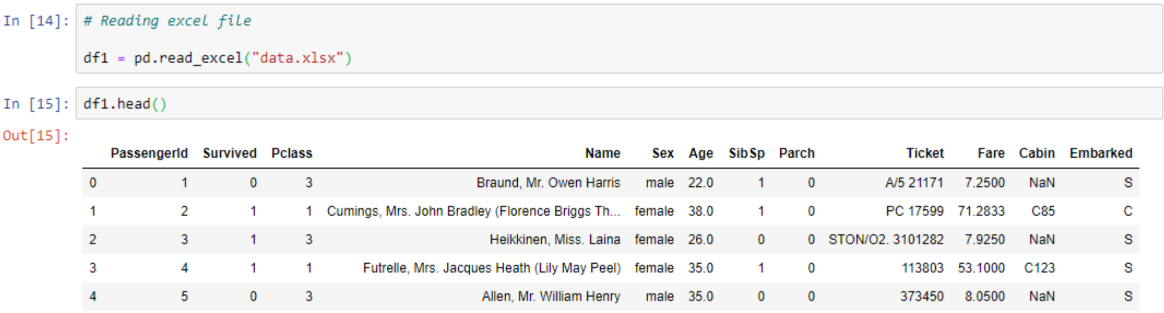


```
In [13]: #selecting multiple columns using names
         df[['Survived','Name']]

Out[13]:
```

	Survived	Name
0	0	Braund, Mr. Owen Harris
1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...
2	1	Heikkinen, Miss. Laina
3	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)
4	0	Allen, Mr. William Henry
...	...	...
886	0	Montvila, Rev. Juozas
887	1	Graham, Miss. Margaret Edith
888	0	Johnston, Miss. Catherine Helen "Carrie"
889	1	Behr, Mr. Karl Howell
890	0	Dooley, Mr. Patrick

891 rows × 2 columns



In [14]:
```
# Reading excel file

df1 = pd.read_excel("data.xlsx")
```

In [15]:
```
df1.head()
```

Out[15]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Observation: In Python, spreadsheets can be read and analysed.

Practice Questions:

1. The file ‘data2.csv’ is the Annual Balance Sheet and contains accumulated accounts and data from 2008 to 2017.
   (i) How many rows and columns are present?
   (ii) View the top 6 rows.
   (iii) What information is being compared in the sheet?

2. For the file ‘data2.csv’,
   (i) Find the count of duplicated data in the dataset.
   (ii) Find the count of duplicated data in the column ‘Institutional_sector_code’.
   (iii) Remove the duplicated data in the column ‘Institutional_sector_code’.

3. For the file ‘data2.csv’,
   (i) Find the count of duplicated data in columns, ‘Institutional_sector_code’ and ‘Status’.
   (ii) Remove this duplicated data and print the number of rows and columns.
   (iii) Print the original data.