	Assignment No. 01 **Semester Spring 2021** **Introduction To Data Science** **Name:** Salman Abdul Rahim **Roll#:** 616	**Due Date : 29-October-2021**

Instructions

Please read the following instructions carefully before solving & submitting assignment:

- **It should be clear that your assignment will not get any credit (zero marks) if:**
- **The assignment is submitted after due date.**
- **The submitted assignment does NOT open or file is corrupted.**
- **The assignment is copied (from other student or copy from handouts or internet).**
- **Source code for Q2 is required**. Name/document your functions appropriately. To make sure that your program can run by the grader, please explicitly import all needed packages .

Uploading instructions

For clarity and simplicity, You are required to Upload/Submit only PDF or MS word file.

Objective

The objective of this assignment is to make you familiar with components of computer and its functionality.

1. **Pandas basics**

   Let df be a pandas DataFrame constructed with the following code:

```
In [62]: data = np.array([0, 7, 3, 6, 2, 8, 5, 9, 4]).reshape(3, -1)
In [63]: df = pd.DataFrame(data, index=['One', 'Two', 'Three'], columns=['a', 'b', 'c'])
```

   What is the output of the following code? (Try to write the output without using python.)

   A. print(df)

```
       a  b  c
One    0  7  3
Two    6  2  8
Three  5  9  4
```

a. df[‘a’]

```
One      0
Two      6
Three    5
```

b. df[‘One’]

```
KeyError: 'One'
```

c. df.loc[‘Two’]

```
a    6
b    2
c    8
```

d. df[:2]

	a	b	c
One	0	7	3
Two	6	2	8

e. df.iloc[:,:2]

	a	b
One	0	7
Two	6	2
Three	5	9

f. list(df.columns)

```
['a', 'b', 'c']
```

g. list(df.index)

```
['One', 'Two', 'Three']
```

h. df[‘b’][‘Two’]

```
2
```

i. list(df.iloc[2, :])

```
[5, 9, 4]
```

j. df.drop('a', axis=1)

	b	c
One	7	3
Two	2	8
Three	9	4

k. df[df.a !=5]

	a	b	c
One	0	7	3
Two	6	2	8

l. list(df.sum(axis=0))

```
[11, 18, 15]
```

m. df.iloc[:, list(df.sum(axis=0) < 17)]

	a	c
One	0	3
Two	6	8
Three	5	4

n. df.sort_values(by='c')

	a	b	c
One	0	7	3
Three	5	9	4
Two	6	2	8

o. df.sort_values(by='Two', axis=1)

	b	a	c
One	7	0	3
Two	2	6	8
Three	9	5	4

p. df.T

	One	Two	Three
a	0	6	5
b	7	2	9
c	3	8	4

q. (df<=2).any(axis=0)

```
a     True
b     True
c    False
```

r. df.applymap(lambda x: x*2-1)

	a	b	c
One	-1	13	5
Two	11	3	15
Three	9	17	7

s. df.apply(lambda x: max(x), axis=1)

```
One      7
Two      8
Three    9
```

2. Multiple Linear Regression

a. What is HDF5 files in Python?

An HDF5 file is a container for two kinds of objects: datasets, which are array-like collections of data, and groups, which are folder-like containers that hold datasets and other groups. The most fundamental thing to remember when using h5py is:
''Groups work like dictionaries, and datasets work like NumPy arrays''.

b. Load data stored in HDF5 format into python using the following statement: hdfstore = pd.HDFStore('hw3q3.h5'). Perform a least square multiple linear regression between the objects x and y in hdfstore (hdfstore['x'] and hdfstore['y']). Report the R-squared and Mean Square Error (MSE) of the regression. Plot the coefficients in a bar chart.

(Code is attached as the .ipynb file)

GOOD LUCK

Marks: 5