AMRITA
VISHWA VIDYAPEETHAM

श्रद्धावान् लभते ज्ञानम्

NAME : P.MOHITH PAVAN KUMAR

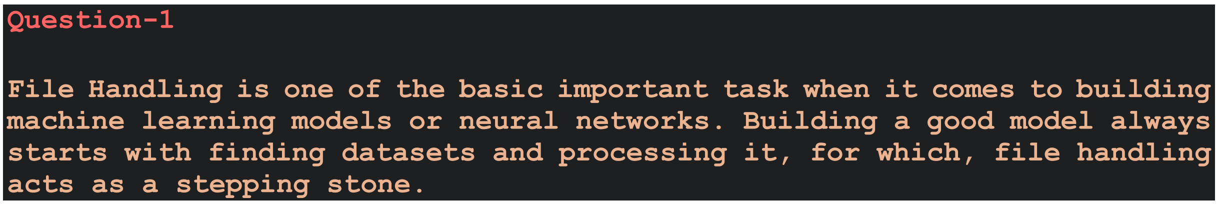
CLASS : CSE-B Group-B

ROLL NO : CH.EN.U4CSE20148

COGNIZANCE TASK 6

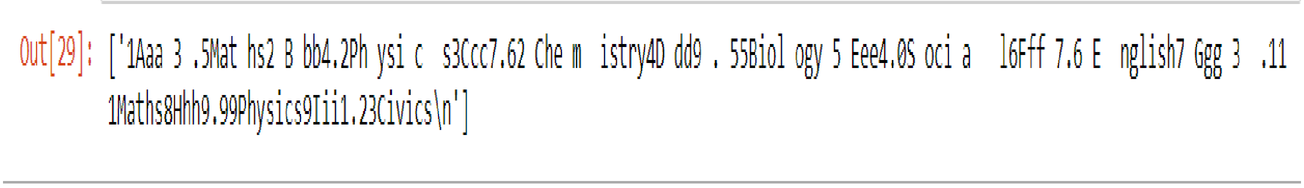
Amrita School of Engineering

Chennai

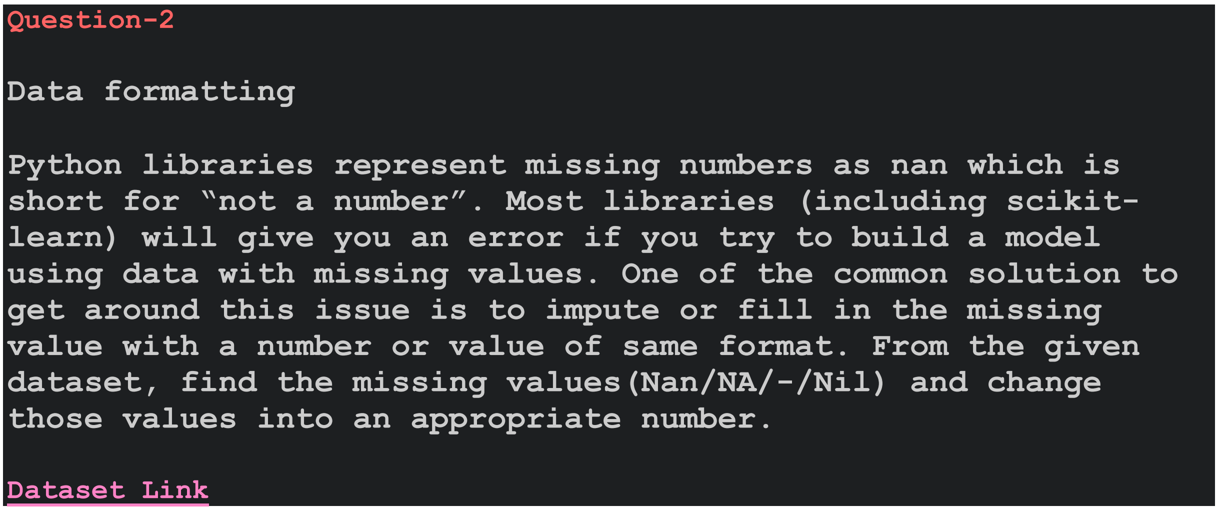


```
Question-1

File Handling is one of the basic important task when it comes to building
machine learning models or neural networks. Building a good model always
starts with finding datasets and processing it, for which, file handling
acts as a stepping stone.
```



```
Out[29]: ['1Aaa 3 .5Mat hs2 B bb4.2Ph ysi c  s3Ccc7.62 Che m  istry4D dd9 . 55Biol ogy 5 Eee4.0S oci a  16Fff 7.6 E  nglish7 Ggg 3  .11
1Maths8Hhh9.99Physics9Iii1.23Civics\n']
```



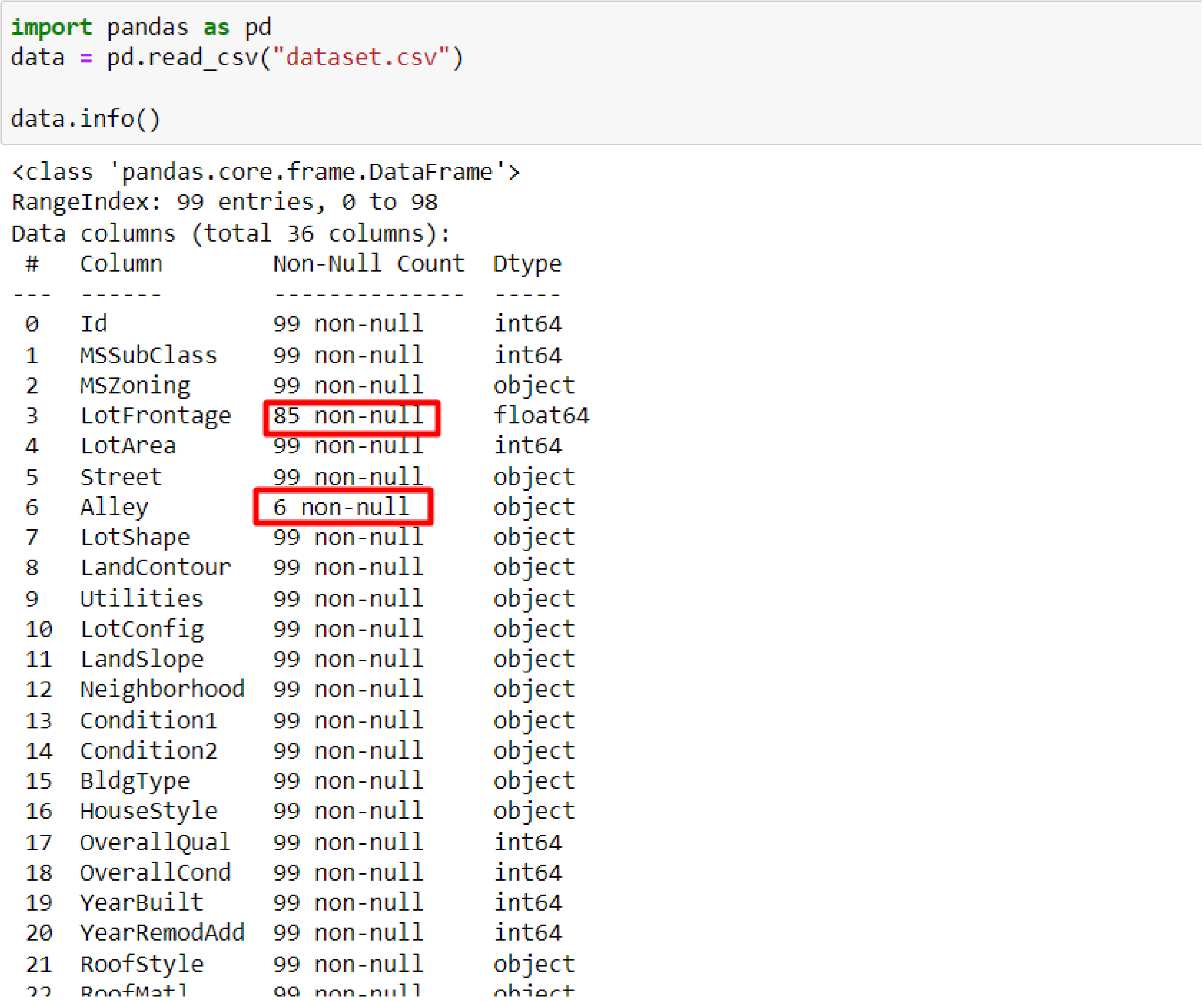
Question-2

Data formatting

Python libraries represent missing numbers as nan which is short for “not a number”. Most libraries (including scikit-learn) will give you an error if you try to build a model using data with missing values. One of the common solution to get around this issue is to impute or fill in the missing value with a number or value of same format. From the given dataset, find the missing values(Nan/NA/-/Nil) and change those values into an appropriate number.

Dataset Link

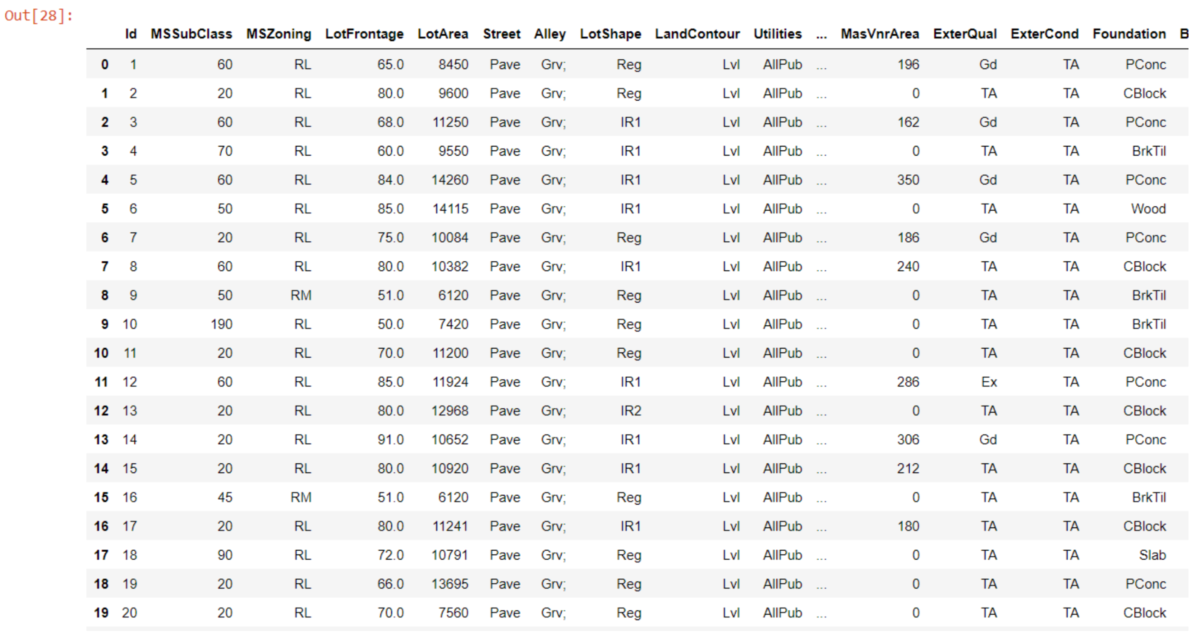
Before we have some null values



```
import pandas as pd
data = pd.read_csv("dataset.csv")

data.info()
```

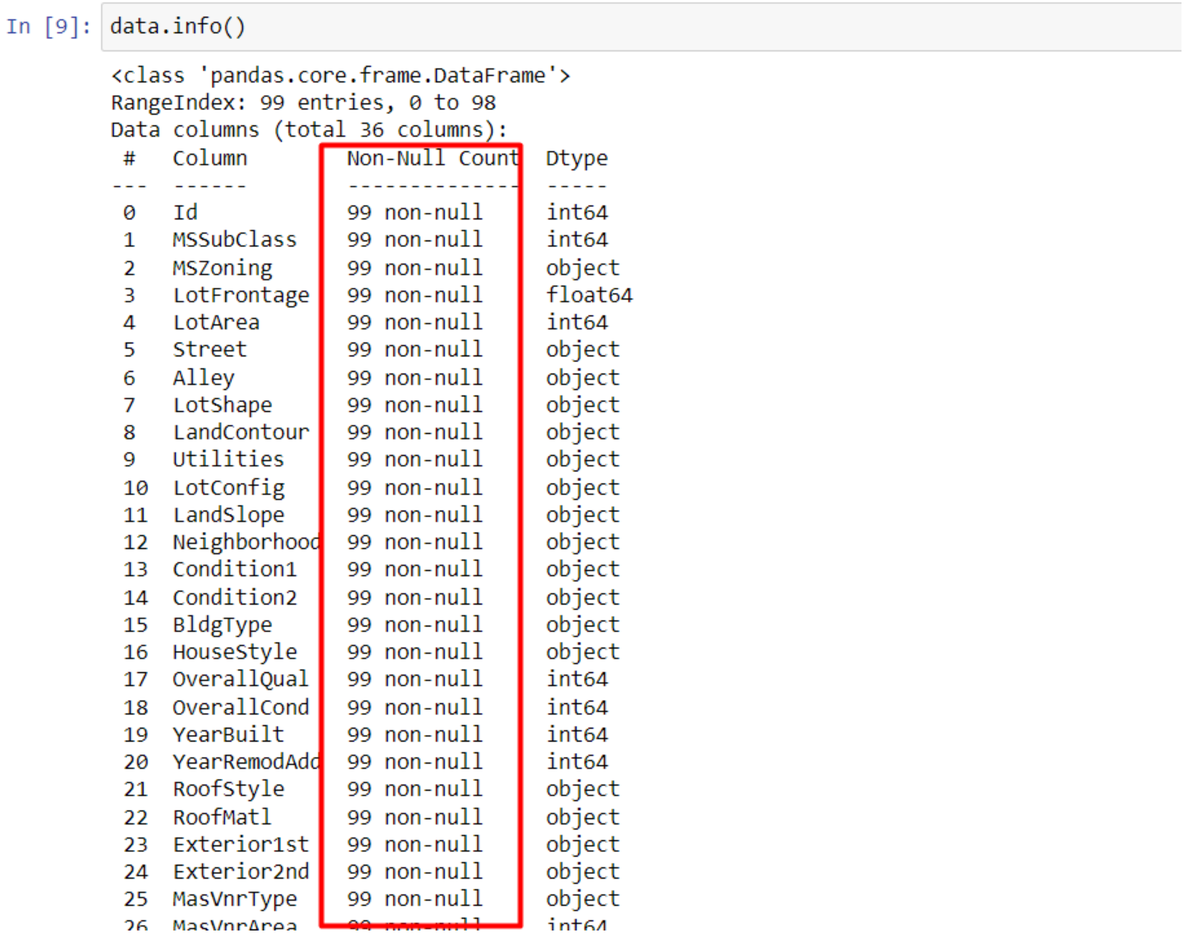
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 36 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Id            99 non-null     int64  
 1   MSSubClass    99 non-null     int64  
 2   MSZoning      99 non-null     object 
 3   LotFrontage   85 non-null     float64
 4   LotArea       99 non-null     int64  
 5   Street        99 non-null     object 
 6   Alley         6 non-null      object 
 7   LotShape      99 non-null     object 
 8   LandContour   99 non-null     object 
 9   Utilities     99 non-null     object 
 10  LotConfig     99 non-null     object 
 11  LandSlope     99 non-null     object 
 12  Neighborhood  99 non-null     object 
 13  Condition1    99 non-null     object 
 14  Condition2    99 non-null     object 
 15  BldgType      99 non-null     object 
 16  HouseStyle    99 non-null     object 
 17  OverallQual   99 non-null     int64  
 18  OverallCond   99 non-null     int64  
 19  YearBuilt     99 non-null     int64  
 20  YearRemodAdd  99 non-null     int64  
 21  RoofStyle     99 non-null     object 
 22  RoofMatl      99 non-null     object 
```



Out[28]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	MasVnrArea	ExterQual	ExterCond	Foundation	B
0	1	60	RL	65.0	8450	Pave	Grv;	Reg	Lvl	AllPub	...	196	Gd	TA	PConc	
1	2	20	RL	80.0	9600	Pave	Grv;	Reg	Lvl	AllPub	...	0	TA	TA	CBlock	
2	3	60	RL	68.0	11250	Pave	Grv;	IR1	Lvl	AllPub	...	162	Gd	TA	PConc	
3	4	70	RL	60.0	9550	Pave	Grv;	IR1	Lvl	AllPub	...	0	TA	TA	BrkTil	
4	5	60	RL	84.0	14260	Pave	Grv;	IR1	Lvl	AllPub	...	350	Gd	TA	PConc	
5	6	50	RL	85.0	14115	Pave	Grv;	IR1	Lvl	AllPub	...	0	TA	TA	Wood	
6	7	20	RL	75.0	10084	Pave	Grv;	Reg	Lvl	AllPub	...	186	Gd	TA	PConc	
7	8	60	RL	80.0	10382	Pave	Grv;	IR1	Lvl	AllPub	...	240	TA	TA	CBlock	
8	9	50	RM	51.0	6120	Pave	Grv;	Reg	Lvl	AllPub	...	0	TA	TA	BrkTil	
9	10	190	RL	50.0	7420	Pave	Grv;	Reg	Lvl	AllPub	...	0	TA	TA	BrkTil	
10	11	20	RL	70.0	11200	Pave	Grv;	Reg	Lvl	AllPub	...	0	TA	TA	CBlock	
11	12	60	RL	85.0	11924	Pave	Grv;	IR1	Lvl	AllPub	...	286	Ex	TA	PConc	
12	13	20	RL	80.0	12968	Pave	Grv;	IR2	Lvl	AllPub	...	0	TA	TA	CBlock	
13	14	20	RL	91.0	10652	Pave	Grv;	IR1	Lvl	AllPub	...	306	Gd	TA	PConc	
14	15	20	RL	80.0	10920	Pave	Grv;	IR1	Lvl	AllPub	...	212	TA	TA	CBlock	
15	16	45	RM	51.0	6120	Pave	Grv;	Reg	Lvl	AllPub	...	0	TA	TA	BrkTil	
16	17	20	RL	80.0	11241	Pave	Grv;	IR1	Lvl	AllPub	...	180	TA	TA	CBlock	
17	18	90	RL	72.0	10791	Pave	Grv;	Reg	Lvl	AllPub	...	0	TA	TA	Slab	
18	19	20	RL	66.0	13695	Pave	Grv;	Reg	Lvl	AllPub	...	0	TA	TA	PConc	
19	20	20	RL	70.0	7560	Pave	Grv;	Reg	Lvl	AllPub	...	0	TA	TA	CBlock	

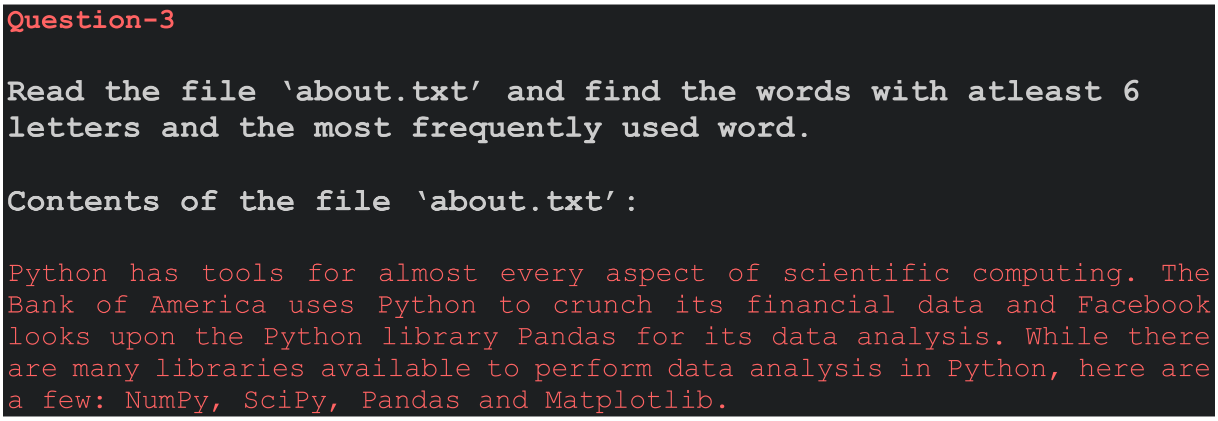
After we filled those with appropriate values



In [9]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 36 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Id            99 non-null     int64  
 1   MSSubClass    99 non-null     int64  
 2   MSZoning      99 non-null     object 
 3   LotFrontage   99 non-null     float64
 4   LotArea       99 non-null     int64  
 5   Street        99 non-null     object 
 6   Alley         99 non-null     object 
 7   LotShape      99 non-null     object 
 8   LandContour   99 non-null     object 
 9   Utilities     99 non-null     object 
 10  LotConfig     99 non-null     object 
 11  LandSlope     99 non-null     object 
 12  Neighborhood  99 non-null     object 
 13  Condition1    99 non-null     object 
 14  Condition2    99 non-null     object 
 15  BldgType      99 non-null     object 
 16  HouseStyle    99 non-null     object 
 17  OverallQual   99 non-null     int64  
 18  OverallCond   99 non-null     int64  
 19  YearBuilt     99 non-null     int64  
 20  YearRemodAdd  99 non-null     int64  
 21  RoofStyle     99 non-null     object 
 22  RoofMatl      99 non-null     object 
 23  Exterior1st   99 non-null     object 
 24  Exterior2nd   99 non-null     object 
 25  MasVnrType    99 non-null     object 
 26  MasVnrArea    99 non-null     int64  
```

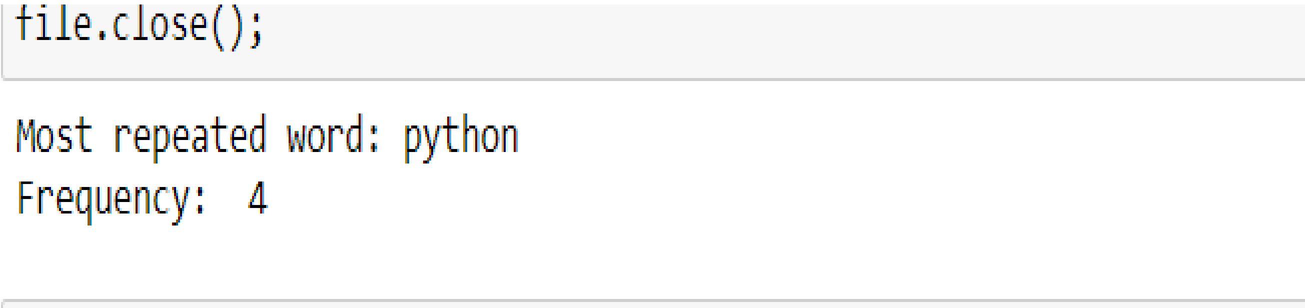


Question-3

Read the file ‘about.txt’ and find the words with atleast 6 letters and the most frequently used word.

Contents of the file ‘about.txt’:

Python has tools for almost every aspect of scientific computing. The Bank of America uses Python to crunch its financial data and Facebook looks upon the Python library Pandas for its data analysis. While there are many libraries available to perform data analysis in Python, here are a few: NumPy, SciPy, Pandas and Matplotlib.



```
file.close();
```

```
Most repeated word: python
Frequency:  4
```