PDS Assignment

1] What do you mean by missing values? Explain the different way to handle the missig value with example? Explain how to deal with missing Data can in pandas.

→ Data can have missing values for an number of a reason such as observation that were not recorded and data corruption. Handling missing data is important as many machine learing alogarithms do not support data with missing values

- In Phython specifically pandas, numpy and skitlearn we mark missing values as nan values with nan value are ingorted from arithmetic operation

- Pandas dataframe provide a function isnull() it returns a new dataframe of same size as calling data it contains only True or false with True at place nan in orginal dataframe and false at other place.

→ Eg

→
```
import pandas as pd
import numpy as np

df = pd.DataFrame(np.random.randn(3,3),
index=['a','c','e'], columns=['one','two','three'])

df = df.reindex(['a','b','c','e'])
Print('nan replace with 0..')
Print(df.fillna())
```

O/P

	one	two	three
a	0.523	0.869	0.725
b	nan	nan	nan
c	0.869	0.354	0.123

2) Explain Dataframe in Pandas ?

- A Pandas frame is a 2 dimensional data structure like a 2-D array or a table with row and colum.

or

- Pandas DataFrame is 2D size-mutable potentially heterog enous tabular ds with lable axes.

- A Data frame is a 2-dimensional data structure i.e data aligned in a tabular fashion in row and colums

- Pandas dataframe consits of three principal component the data, row and colum

Eg

```
import pandas as pd

data {"calaries": [420, 380, 390],
      duration": [50, 45, 40]}

df = pd.DataFrame(data)
print(df)
```

O/P

	calaries	duration
0	420	50
1	380	40
2	390	45

3) compare numpy vs Pandas

Pandas	numpy
• Pandas is mostly use for dataanyalsis	• numpy is mostly work with numerical values
• It consumes more memory	• It consumes less memory
• Indexing of pandas series is very slow.	• Indexing of numpy array is fast.
• Pandas have 2d labal, object called Dataframe.	• numpy provides a nutidemisonal arrays.
• Dataframe and series are most povorful tools for pandas.	• Array is a most povorful for numpy.
• It is used in higher industry	• It is used in lower industry.

4) Explain TF-IDF technique

→ TF-IDF stands for Term frequency Inverse data frequency of word.

- It can be defined as calculation of how relvant a word in series or dataset.
- The meaning increase proportionally to the number of time in text a word frequency in the dataset.

- Time frequency (TF) - It gives us frequency of word in each document.

- Inverse Data frequency (IDF) - IDF is used to calculate the weight of rare words across all document.

→ Eg
```
import numpy as num
from sklearn.feature_extraction.text import
string = ['Geeks geeks'] * 5
tfidf = Tfidf vectorizer()
result = tfidf.fit_transform(string)
result =
print (tfidf.vocabulary)
print (result)
```

O/P

word indexs		ti-idf value
{'geeks': 0	(0,0)	1.0
	(1,0)	1.0
	(2,0)	1.0
	(3,0)	1.0
	(4,0)	1.0

5] Difference between rand() and randn() function in numpy

Rand()	Randn()
- generate uniform random number	- generate a gaussian random number
- Rand() have one random number	- A Randn() have n random numbers
- A Rand() have a positive random	- A Randn() have negative random