Rahul Gohil 2002801070096

# Assignment - 2

Q-1 We have student data set with attributes Name, Enrollment NO, semester, CGPI, Result status

→ First we have to import Pandas module & the dataset.

```
import Pandas as pd.
dt = pd.read.CSV("sample.CSV")
```

-> There are mainly 4 technique to handle missing data.

1) Delete the rows with missing values.

```
# deleting rows if all vals are
missing
df.dropna(how = "all")
```

2) Replace the missing vals with mean or median

```
df[['CGPI']].fillna(value = df['CGPI']
    .mean(), inplace = true)

df['Semester'].fillna(value = df
    ['Semester'].median(),
    inplace = True)
```

3) Imputation method for categorical columns.

```
df['name'].fillna("UnKnown",
    inplace = True).
```

vision

Rahul Gohil 2002801070 96

4) If some data has longitudinal behaviour, then it will make sense to use the last valid observation

```
df['semester'].fillna(method='ffill')
```

Q-2

Random sampling refers to randomly selecting rows from the dataframes.

→ sample() fun$^{n}$ is used to get sample data from the provided data frame.

→ Example :-

```
import pandas as pd
df = pd.read_csv("Data.csv")
# To get 1 row randomly from df
df.sample()
# To get 3 rows of sample data.
df.sample(n=3)
#To get 70% random data.
df.sample(frac=0.7)
```

Rahul Gohil 20028010 7096

# when we want no. of rows of sample that may exceed available rows.
df.sample(n=100, replace = True).

Q-3 rand & randn in numpy :

→ randn() : It generates an array of specified shape & fills it with random float value as per standard ~~not~~ normal distribution.

→ Mean of all val is 0 & Variance is 1

→ rand() : It generates an array of specified shape & fills it with random val. ,
It doesn't fill value as per Std. normal destribution.

⇒ Join and Merge in Panday :-

Join() : It combines 2 data frames on the bases of their indices.

VISION

Rahul Gohil 2002801070 96

Merge() : It is more versatile & allows us to specify columns beside the index to join on for both dataframes.

Q-4 Attributes of the dataset : Name, En. no, Sem, CGPI, Result Status.

```
import Pandas as pd
df = pd.read_csv("sample.csv")

#Grouping the data by semester
& counting
# freq. for each sem
grp = df.groupby("sem").count()
Print(grp).
```

Q-5 n-grams : They are continuous seq. of words or symbols or tokens in a document.

- Bag of words simply refers to a matrix whole rows are docs & columns are words.
- whereas, IF gives the freq. of the word in each doc. & IDF calculate the weight of rare words.

vision

Rahul Gohil 200280107096

Q-6 Regular expressions are specially encoded text strings used as patterns for matching sets of strings.

→ Conditions for valid GTU enrollment no. :-
- length should be 12
- Contains only no.

→ Code for validating enrollment no :-

```
import re
test = input()
# for second conditions,

def check1(test):
    if re.match("^[0-9]", test)
        return false.
else return true.

# For first condition

def check2(test):
    if (test.. len() == 12)
        return true
else return false.
```

```
if (check1(text) and check2(text)):
    print ("valid enrollment no.")
else
    print ("Invalid enrollment no.")
```