1. //import all python libraries

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

2. www.kaggle.com

3. //load dataset into pandas data frame

```
df=pd.read_csv("abc.csv")
   print(df)
```

4. Data preprocessing:-

4.1 Describe Function:

```
df.describe()//to get some initial statistics
```

4.2Check dimension of data frame

```
df.shape // will display no. row and col

df.head(2)//first 2 row
df.tail(1)//last 1 row
```

4.3types of variable

```
type(1)//int
type("abc")//str
type(4.2)//float
```

4.4 check for missing values in data frame

//Checking for missing values

```
df.isnull()


df.isnull().sum()


df.isnull().sum().sum()
```

//Fill null value with different value

```
df2=df.fillna(value = 0)
df2
```

//Fill null value with previous row value

```
df4=df.fillna(method='pad')
df4
```

//Fill null value with next (Backword) row value

```
df5=df.fillna(method='bfill')
df5
```

//Fill null value with previous column value

```
df6=df.fillna(method='pad',axis=1)
df6
```

//Fill null value with next (Backword) Column value

```
df7=df.fillna(method='bfill',axis=1)
df7
```

//filling with different values in Null in different column

```
Df7=df.fillna({'Roll_no':'abcd'})
Df7
```

//filling null value with the mean/max/min value of a column

```
Df8=df.fillna(value=df['Roll_no'].mean())
Df8

Df9=df.fillna(value=df['Roll_no'].min())
Df9

Df10=df.fillna(value=df['Roll_no'].max())
Df10
```

//Drop such missing value use dropna() function

```
df5=df.dropna()
df5
```

//replace Null value

```
import numpy as np
df5=df.replace(to_replace=np.nan,value=123)
df5
```

5. data formatting and data normalization in python

```
//dtype() to check data type
//astype() to change Data type
```

```
df.dtypes
//now change marks col data type from int to float

df['marks']=df['marks'].astype(float)
df['marks'].dtypes



//now change marks col data type from float  toint



df['marks']=df['marks'].round(0).astype(int)
df['marks'].dtypes
```

//pd.to-numeric function

```
df['Roll_no']=pd.to_numeric(df['Roll_no'].round(0), downcast='integer')
df['Roll_no'].dtypes


//if we want to convert into integer


df['Roll_no']=pd.to_numeric(df['Roll_no']).astype('Int32')
df['Roll_no'].dtypes
```

6. how to convert categorical variables into quantitative variables in python

```
import pandas as pd
import numpy as np
iris=pd.read_csv("iris.csv")
print(df)
```

```
iris['code']=pd.factorize(iris.Species)[0]
iris.Species.value_counts()
```

output

```
setosa        50
versicolor    50
virginica     50
   Name: Species, dtype: int64
```