TOPIC:**DATA QUALITY ISSUE ON ADDRESS**
NAME:S.Hari krishnan-Intern

PROBLEM STATEMENT:

The quality of data is determined by factors such as accuracy, completeness, reliability, relevance and how up to date it is. As data has become more intricately linked with the operations of organizations, the emphasis on data quality has gained greater attention.Using Machine learning,the good data and bad data are identified and separated.Approximately,to some extent bad data are removed.

Why data quality is important?

Poor-quality data is often pegged as the source of inaccurate reporting and ill-conceived strategies in a variety of companies, and some have attempted to quantify the damage done.

DEVELOPMENT ENVIRONMENT USED:

GOOGLE COLABORATORY -python3
DATASET USED-OWN

PROCESS:

1.DATA EXPLORATION:

data=pd.read_csv("https://raw.githubusercontent.com/harkrish/datasetaddress/master/address2.csv")
data.head(6)

	Address	Unnamed: 1	Unnamed: 2
0	ramapuram	NaN	NaN
1	#3, Sr no 33/6, saifitness building, near icc...	NaN	NaN
2	1027 sector 28 ground floor faridabad haryana	NaN	NaN
3	106b U&V block shalimar bagh delhi	NaN	NaN
4	1077,sector-4/a,bokaro steel city, jharkhand,8...	NaN	NaN
5	NaN	NaN	NaN

2.IDENTIFY MISSING VALUES IN DATASET:

```
totalcells=np.product(x.shape)
missingCount = x.isnull().sum()
totalMissing = missingCount.sum()
print(" dataset contains", round(((totalMissing/totalcells) * 100), 2),
"%", "missing values.")
```

```
dataset contains 18.18 % missing values.
```

3.TO FILL NA/NAN VALUES BY SPECIFIC METHOD:

```
inpute = data.fillna(method='ffill', axis=0).fillna("0")
inpute.head(6)
```

	Address	Unnamed: 1	Unnamed: 2
0	ramapuram	0	0
1	#3, Sr no 33/6, saifitness building, near icc...	0	0
2	1027 sector 28 ground floor faridabad haryana	0	0
3	106b U&V block shalimar bagh delhi	0	0
4	1077,sector-4/a,bokaro steel city, jharkhand,8...	0	0
5	1077,sector-4/a,bokaro steel city, jharkhand,8...	0	0

4.CLEANING ADDRESS DATA:

```
address_cln = []
 addrs = inpute.Address
```

STEP 1: Convert the address to lower case:

```
addr = str(addr)
addr = addr.lower()
```

STEP 2: Adjust the commas

1.Counting the number of comma present

```
commas = addr.count(',')
```

2.Checking if there is a space after commas (if not include a space)

```
if commas > 0:
    indx = 0
for i in range(0,commas):
    indx = addr.find(',', indx) + 1
    if addr[indx] == " ":
        continue
    else:
        addr = addr[0:indx] + ' ' + addr[indx:]
```

STEP 3: Remove full stops

```
addr = addr.replace('.', ' ')
```

STEP 4: Remove - with "

```
addr = addr.replace('-', ' ')
```

STEP 5: Check if the string starts with single/double letters followed by a string containing numbers. Join them together

```
firstword = addr[:addr.find(" ")]
nextword  = addr[addr.find(" ")+1:addr.find(" ",addr.find(" ")+1)]

if (len(firstword)) <= 2 and any(char.isdigit() for char in addr):
    addr = firstword + nextword + addr[addr.find(' ', addr.find(' ')+1):]
```

STEP 6:Remove extra white spaces

```
addr = addr.strip()
```

STEP 7: Spot the first '/' and remove the pace immediately before and after it (if any)

```
slash = addr.find('/')
if addr[slash+1] == ' ':
    addr = addr[:slash+1] + addr[slash+2:]
if addr[slash-1] == ' ':
    addr = addr[:slash-1] + addr[slash:]
```

STEP 8: Cleaning brackets and its contents

```
if (addr.find('(') != -1) and (addr.find(')',addr.find('(')) != -1):
    addr = addr[:addr.find('(')] + addr[addr.find(')',addr.find('('))+1:]
```

STEP 9: Append data

```
address_cln.append(addr)
```

STEP 10:PRINTING ORIGINAL DATA:

```
                                                Address  ... Unnamed: 4
0                                              ramapuram  ...          0
1    #3, Sr no 33/6, saifitness building, near icc...  ...          0
2       1027 sector 28 ground floor faridabad haryana  ...          0
3                   106b U&V block shalimar bagh delhi  ...          0
4   1077,sector-4/a,bokaro steel city, jharkhand,8...  ...          0
5   1077,sector-4/a,bokaro steel city, jharkhand,8...  ...          0
6   No.67,MUNUSAMY SALAI,K.K NAGAR WEST,CHENNAI-60...  ...          0
7              No.54 shivam apt,T.NAGAR,Chennai-600079  ...          0
8         Vijay apartments,No 56,Naga street,vt nag...  ...          0
9         Vijay apartments,No 56,Naga street,vt nag...  ...          0
10   No 43   , S And P residency,Madhu st,vellore-...  ...          0
```

STEP 10: GOOD DATA OUTPUT:

```
for i in range(len(address_cln)):
    print(address_cln[i])
```

```
ramapuram
3,sr no 33/6, saifitness building, near icc
1027 sector 28 ground floor faridabad haryana
106b u&v block shalimar bagh delhi
1077, sector 4/a, bokaro steel city, jharkhand, 8
1077, sector 4/a, bokaro steel city, jharkhand, 8
no67, munusamy salai, k k nagar west, chennai 600078
no54 shivam apt, t nagar, chennai 600079
vijay apartments, no 56, naga street, vt nagar, madurai 600046
vijay apartments, no 56, naga street, vt nagar, madurai 600046
no43   , s and p residency, madhu st, vellore 780043
```