Kasra Samadi 993623030

HW1 DataMining

Tasks :

1- Read data set-1.csv and find missing data

2- Fill missing data in Online_Gaming column with N value

3- Data reduction in Online_Shopping column

4- Sampleing

5- Inconsistent data

6- Reduction of attributes

Task 1 And 2

Import libraries

In [1]:
```
import numpy as np
import pandas as pd
```

Reading data set-1.csv

In [2]:
```
df = pd.read_csv("data set-1.csv")
```

Missing datas are marked with NaN

In [3]:
```
df
```

Out[3]:

	Gender	Race	Birth_Year	Marital_Status	Years_on_Internet	Hours_Per_Day	Preferred_B
0	M	White	1972	M	8	1	
1	M	Hispanic	1981	S	14	2	(
2	F	African American	1977	S	6	2	
3	F	White	1961	D	8	6	
4	M	White	1954	M	2	3	Internet E
5	M	African American	1982	D	15	4	Internet E
6	M	African American	1981	D	11	2	
7	M	White	1977	S	3	3	Internet E
8	F	African American	1969	M	6	2	
9	M	White	1987	S	12	1	
10	F	Hispanic	1959	D	12	5	(

This dataFrame has 11 rows and 15 columns

In [4]:
```
df.shape
```

Out[4]:
```
(11, 15)
```

Find columns names and their types

In [5]:
```
df.dtypes
```

Out[5]:
```
Gender                    object
Race                      object
Birth_Year                 int64
Marital_Status            object
Years_on_Internet          int64
Hours_Per_Day              int64
Preferred_Browser         object
Preferred_Search_Engine   object
Preferred_Email           object
Read_News                 object
Online_Shopping           object
Online_Gaming             object
Facebook                  object
Twitter                   object
Other_Social_Network      object
dtype: object
```

Find the number of missing values for each columns

In Online_Gaming column we have 3 missing data

In [6]:
```
df.isna().sum()
```

Out[6]:
```
Gender                    0
Race                      0
Birth_Year                0
Marital_Status            0
Years_on_Internet         0
Hours_Per_Day             0
Preferred_Browser         0
Preferred_Search_Engine   0
Preferred_Email           0
Read_News                 1
Online_Shopping           2
Online_Gaming             3
Facebook                  0
Twitter                   0
Other_Social_Network      7
dtype: int64
```

Replace NaN values (Missing data) Of Online_Gaming column with N

In [7]:
```
df["Online_Gaming"].fillna("N", inplace=True)
```

Now we dont have any missing values in Online_Gaming

In [8]:
```
df.isna().sum()
```

Out[8]:
```
Gender                   0
Race                     0
Birth_Year               0
Marital_Status           0
Years_on_Internet        0
Hours_Per_Day            0
Preferred_Browser        0
Preferred_Search_Engine  0
Preferred_Email          0
Read_News                1
Online_Shopping          2
Online_Gaming            0
Facebook                 0
Twitter                  0
Other_Social_Network     7
dtype: int64
```

We have 9 N values and 2 Y values in Online_Gaming column

In [9]:
```
df['Online_Gaming'].value_counts()
```

Out[9]:
```
N    9
Y    2
Name: Online_Gaming, dtype: int64
```

Task 3 : Data reduction in Online_Shopping column

We have 2 missing values in Online_Shopping column

In [10]:
```
df.isna().sum()
```

Out[10]:
```
Gender                    0
Race                      0
Birth_Year                0
Marital_Status            0
Years_on_Internet         0
Hours_Per_Day             0
Preferred_Browser         0
Preferred_Search_Engine   0
Preferred_Email           0
Read_News                 1
Online_Shopping           2
Online_Gaming             0
Facebook                  0
Twitter                   0
Other_Social_Network      7
dtype: int64
```

This dataFrame has 11 rows and 15 columns

In [11]:
```
df.shape
```

Out[11]:
```
(11, 15)
```

We Drop Null values in Online_Shopping Column

In [12]:
```
dropped_Null_df = df.dropna(subset=["Online_Shopping"])
```

Now dropped_Null_df has 9 rows and 15 columns

In [13]:
```
dropped_Null_df.shape
```

Out[13]:
```
(9, 15)
```

we dont have any missing values in Online_Shopping

In [14]:
```
dropped_Null_df.isna().sum()
```

Out[14]:
```
Gender                    0
Race                      0
Birth_Year                0
Marital_Status            0
Years_on_Internet         0
Hours_Per_Day             0
Preferred_Browser         0
Preferred_Search_Engine   0
Preferred_Email           0
Read_News                 1
Online_Shopping           0
Online_Gaming             0
Facebook                  0
Twitter                   0
Other_Social_Network      7
dtype: int64
```

Task 4 : Sampleing

Sample ratio = 0.5

In [15]:
```
sampled_df = dropped_Null_df.sample(frac=0.5)
```

sampled_df has 4 rows and 15 columns

In [16]:
```
sampled_df.shape
```

Out[16]:
```
(4, 15)
```

Task 5 : Inconsistent data

df dataFrame is without sampling and removing null values in Online_Shopping column

In [17]:
```
df.shape
```

Out[17]:
```
(11, 15)
```

In [18]:
```
df
```

Out[18]:

	Gender	Race	Birth_Year	Marital_Status	Years_on_Internet	Hours_Per_Day	Preferred_B
0	M	White	1972	M	8	1	
1	M	Hispanic	1981	S	14	2	(
2	F	African American	1977	S	6	2	
3	F	White	1961	D	8	6	
4	M	White	1954	M	2	3	Internet E
5	M	African American	1982	D	15	4	Internet E
6	M	African American	1981	D	11	2	
7	M	White	1977	S	3	3	Internet E
8	F	African American	1969	M	6	2	
9	M	White	1987	S	12	1	
10	F	Hispanic	1959	D	12	5	(

We have 8 N values , 2 Y values and one 99 value in Twitter column

In [19]:
```
df['Twitter'].value_counts()
```

Out[19]:
```
N     8
Y     2
99    1
Name: Twitter, dtype: int64
```

Replace 99 value into N value in Twitter column

In [20]:
```
df['Twitter'] = df['Twitter'].replace('99', 'N')
```

We have 9 N values and 2 Y values in Twitter column

There is no 99 value in the Twitter column

In [21]:
```
df['Twitter'].value_counts()
```

Out[21]:
```
N    9
Y    2
Name: Twitter, dtype: int64
```

Task 6 : Reduction of attributes

We want to select Birth_Year, Gender, Marital_Status, Race, Years_on_Internet columns

In [22]:
```
df.shape
```

Out[22]:
```
(11, 15)
```

Select Birth_Year, Gender, Marital_Status, Race, Years_on_Internet columns

In [23]:
```
dropped_attributes_df = df[['Birth_Year', 'Gender', 'Marital_Status', 'Race',
```

dropped_attributes_df has 11 rows and 5 columns

In [24]:
```
dropped_attributes_df.shape
```

Out[24]:
```
(11, 5)
```

In [25]:
```
dropped_attributes_df
```

Out[25]:

	Birth_Year	Gender	Marital_Status	Race	Years_on_Internet
0	1972	M	M	White	8
1	1981	M	S	Hispanic	14
2	1977	F	S	African American	6
3	1961	F	D	White	8
4	1954	M	M	White	2
5	1982	M	D	African American	15
6	1981	M	D	African American	11
7	1977	M	S	White	3
8	1969	F	M	African American	6
9	1987	M	S	White	12
10	1959	F	D	Hispanic	12