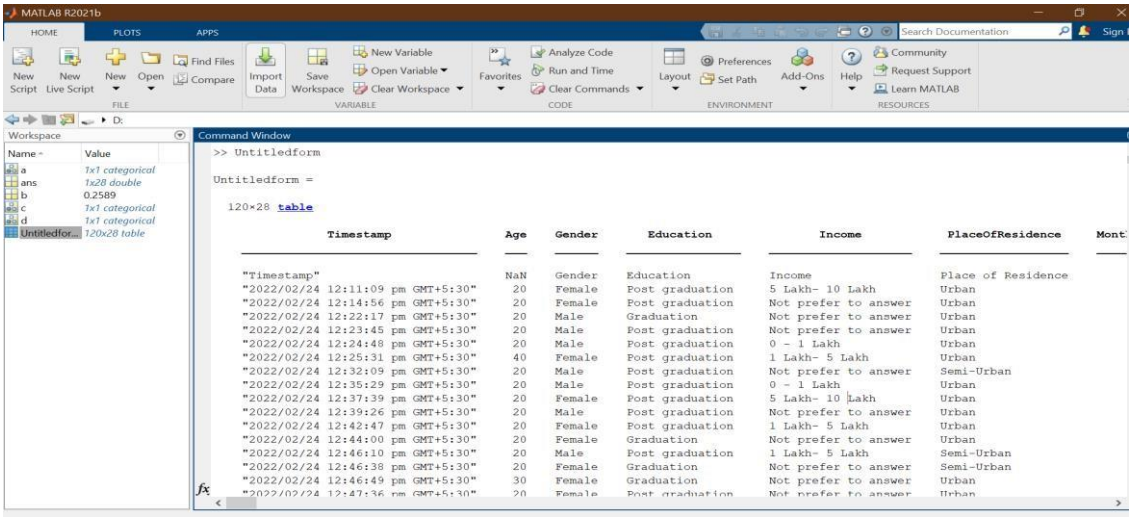
##DATA CLEANING

#Firstly, import dataset in MATLAB using MATLAB importing tool.

>> Untitledform

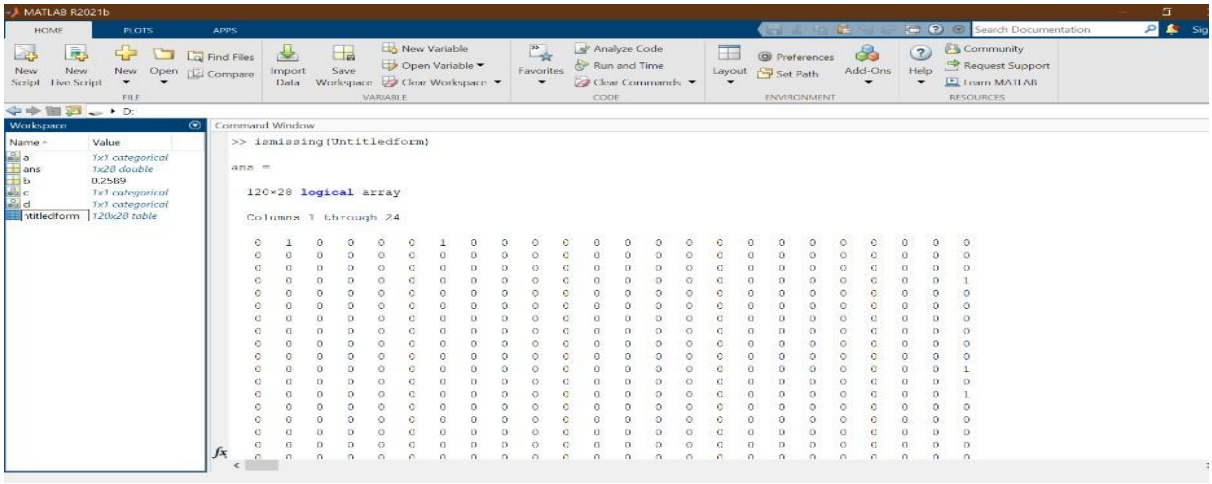
Untitledform =



Incase of numerical data null values are shown as NaN while in categorical data it is shown as <undefined>

Checking, is there any null values are present in the dataset?

For this ismissing command is used which shown in the picture. Here, 1 indicates that there is missing value in that particular column or row.



So, for big data it is difficult to find missing values of columns.

In this case we use below command;

>> sum(ismissing(Untitledform)) ans

=

Columns 1 through 16

0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0

Columns 17 through 28

0 0 0 0 0 0 0 27 0 0 91 0

Here, we can see in 2nd ,7th column there is 1 value is missing & in 24th, 27th column 27 and 91 values are missing respectively.

handling missing values

Basically, data cleaning is done in 2 ways,

1. By removing those columns & rows by using rmmissing command
2. By using mean & mode feature according to their numerical & categorical data respectively

(Reason for using mode feature in categorical data, because mode is that value which has maximum frequency.) >>
d=mode(Untitledform.IfYesWhatLongTermDiseasesDoYouHaveYouCanSelectMoreThanOneOption)

d =

categorical

NA

>>
Untitledform.IfYesWhatLongTermDiseasesDoYouHaveYouCanSelectMoreThanOneOption=fillmissing(Untitledform.IfYesWhatLongTermDiseasesDoYouHaveYouCanSelectMoreThanOneOption,'constant',cellstr(d));

>> Untitledform

Untitledform =

Here, 2nd criteria is used for handling missing values to avoid removing any rows or column.

To replace missing values in original data fillmissing command is used

MATLAB commands are as follows; >>

a=mean(Untitledform.Age,'omitnan')

a =

22.6050

>> Untitledform.Age=fillmissing(Untitledform.Age,'constant',a);

```
>> b=mean(Untitledform.MonthlyExpenditureOnOTCMedicines,'omitnan')


b =

 0.2589

>>Untitledform.MonthlyExpenditureOnOTCMedicines=fillmissing(Untitledform.MonthlyExpenditure
OnOTCMedicines,'constant',b);

>> c=mode(Untitledform.WhichOfTheFollowingSideEffectDoYouSufferYouCanSelectMoreThanOne)

c =  categorical

Sleepiness

>>
Untitledform.WhichOfTheFollowingSideEffectDoYouSufferYouCanSelectMoreThanOne=fillmissing(U
ntitledform.WhichOfTheFollowingSideEffectDoYouSufferYouCanSelectMoreThanOne,'constant',cells
tr(c));

>>
d=mode(Untitledform.IfYesWhatLongTermDiseasesDoYouHaveYouCanSelectMoreThanOneOption)


d =

categorical

NA

>>Untitledform.IfYesWhatLongTermDiseasesDoYouHaveYouCanSelectMoreThanOneOption=fillmissi
ng(Untitledform.IfYesWhatLongTermDiseasesDoYouHaveYouCanSelectMoreThanOneOption,'consta
nt',cellstr(d));
```

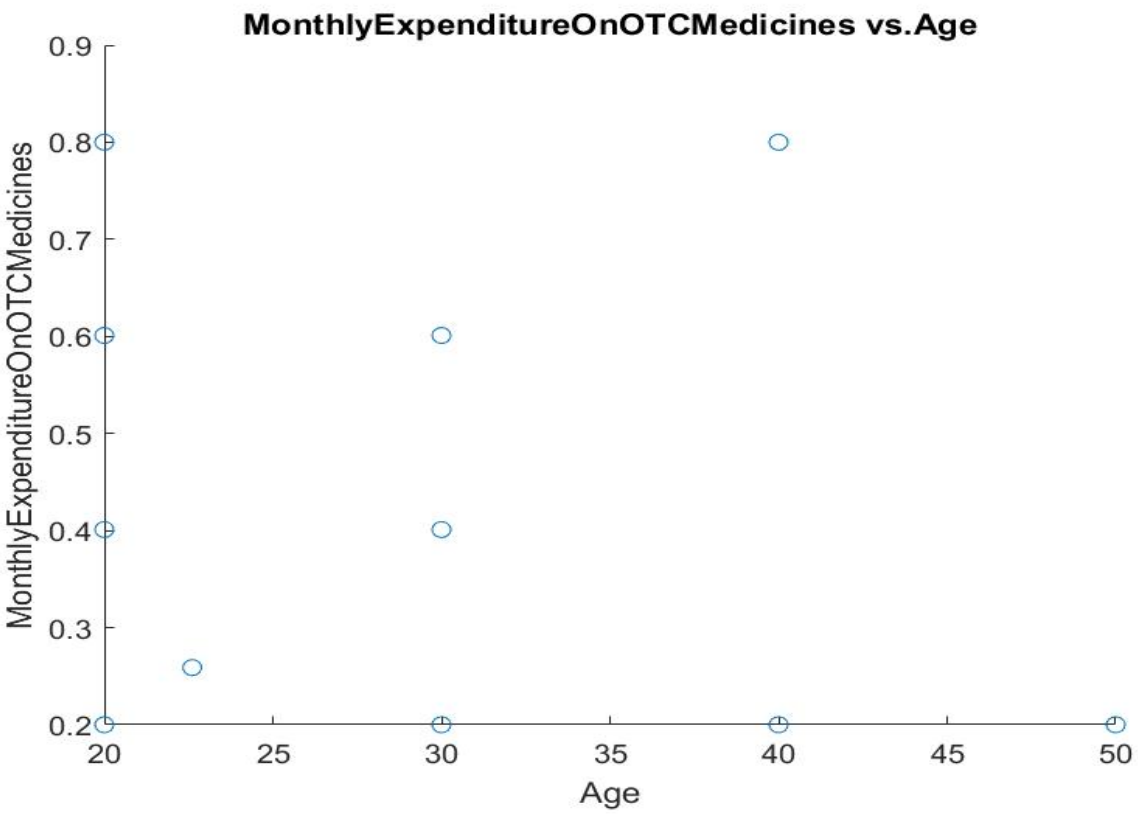
Lastly, run the original table code;

```
>> Untitledform

Untitledform =
```

Lastly, we get data which has no missing values involved in it.

```
##VISUALIZATION

# scatter plot

>>plot=scatter(Untitledform.Age,Untitledform.Monthly expenditure on OTC medicines.)

>> xlabel('Age')

>> ylabel('MonthlyExpenditureOnOTCMedicines')

>> title('MonthlyExpenditureOnOTCMedicines vs.Age')
```

>> legend



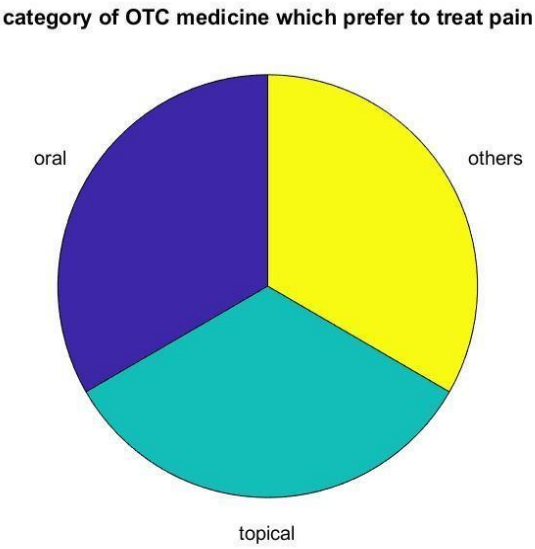
#pie chart

>> x=categorical({'Oral (tablet, syrup, capsule, powder, etc taken internally)','Topical (ointments, cream, liquids that are applied to the skin)','Others (such as eye drop and surgical dressings)'}); >> explode={};

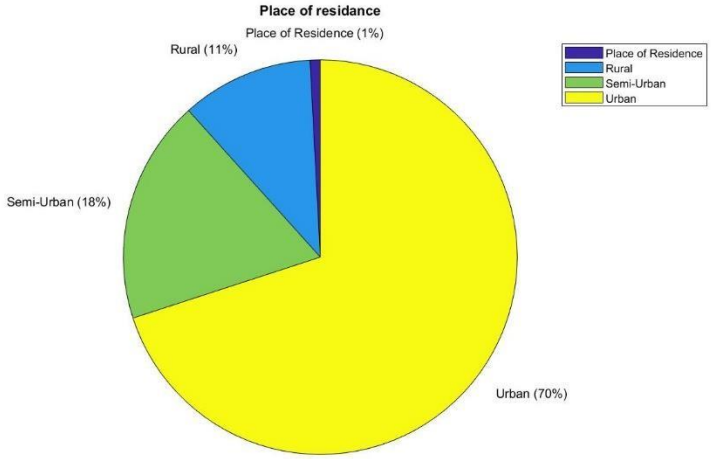
>> label={'oral','topical','others'};

>> pie(x,explode,label)

>> title("category of OTC medicine which prefer to treat pain")

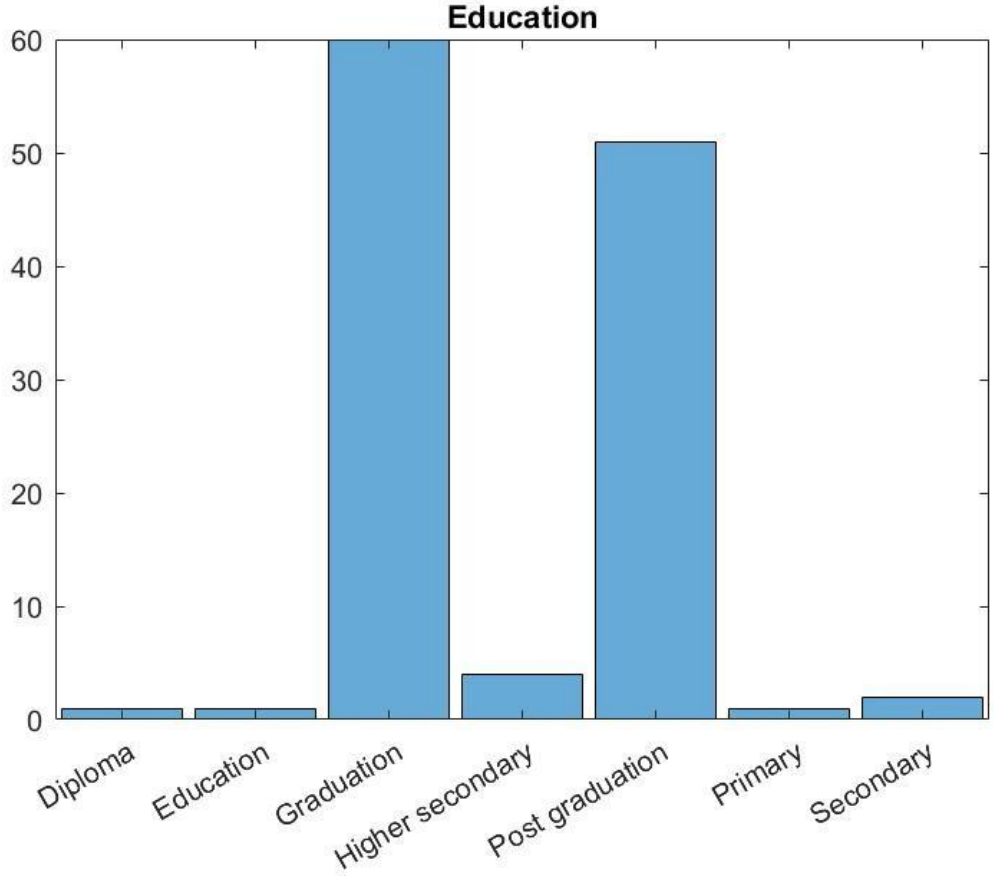


>>pie(Untitledform.PlaceOfResidence)



histogram

>>histogram(Untitledform.Education)


Education
60
50
40
30
20
10
0
Diploma
Education
Graduation
Higher secondary
Post graduation
Primary
Secondary