Contents

Importing Recidivism data set%%

Source of data: https://www.openml.org/d/42193 %

```
clear all;
clc;
close all;
```

```
data = readtable("compas.csv");
data;
```

```
Warning: Column headers from the file were modified to make them valid MATLAB
identifiers before creating variable names for the table. The original column
headers are saved in the VariableDescriptions property.
Set 'VariableNamingRule' to 'preserve' to use the original column headers as
table variable names.
```

On viewing data table it is found that the target column or % or the response variable is in the middle of the table. % So, the response variable is moved to the last column. %

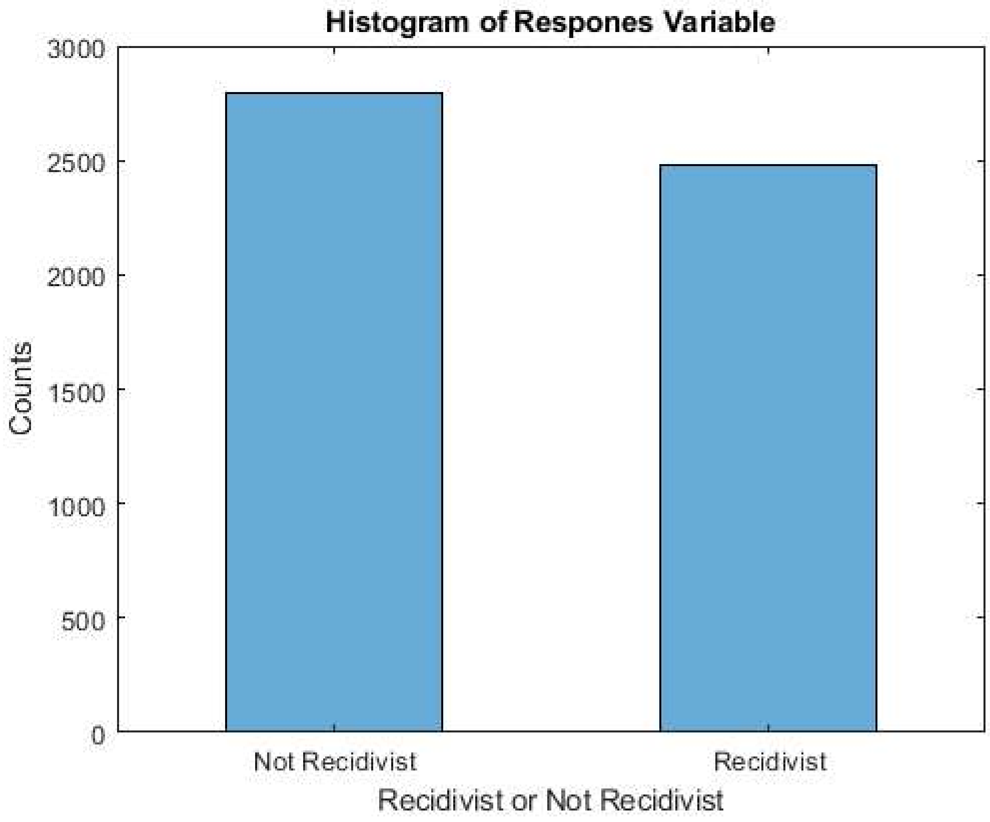
```
data1 = movevars(data,"two_year_recid","After","c_charge_degree_M");
```

Convert reponse variable into categorical data type for further use. %

```
targetcol = categorical(data1.two_year_recid);
```

Plot histogram of response variable to see the proportions of outputs. %

```
figure
histogram(targetcol,'BarWidth',0.5)
title('Histogram of Respones Variable')
xlabel('Recidivist or Not Recidivist')
ylabel('Counts')
xticklabels({'Not Recidivist', 'Recidivist'})
```

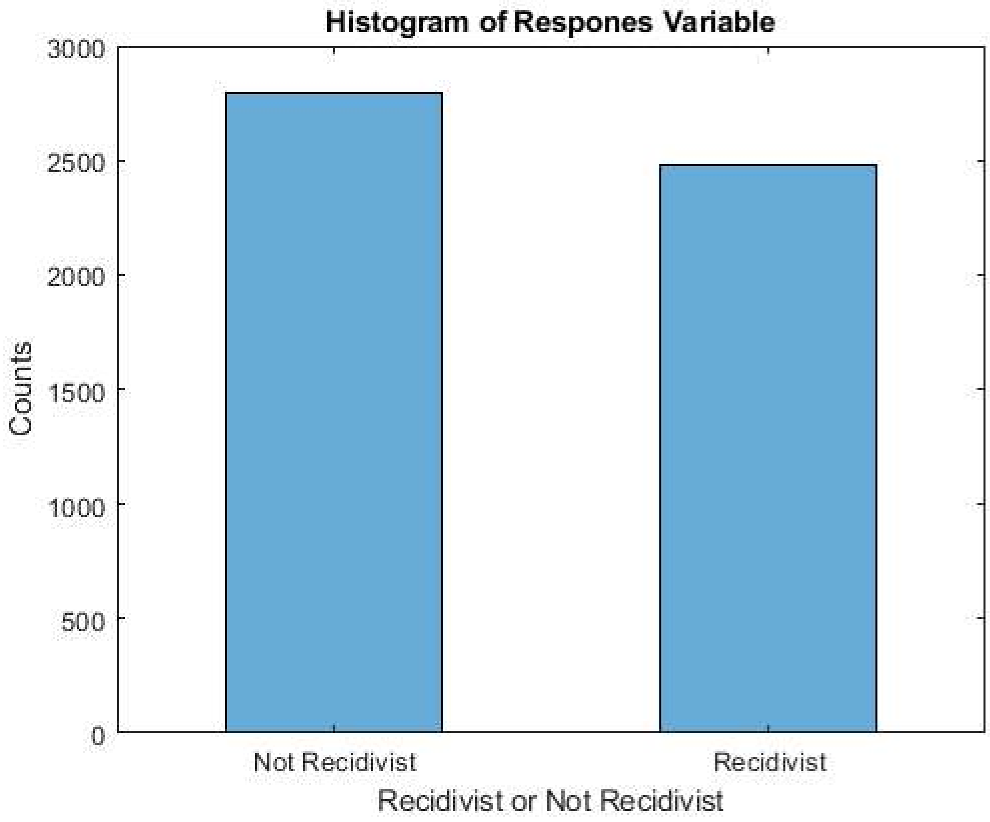


Convert various categorical variables into categorical data type for % further use. %

```
sex = categorical(data1.sex);
agebelow25 = categorical(data1.age_cat_Lessthan25);
age25to45 = categorical(data1.age_cat_25_45);
ageabove45 = categorical(data1.age_cat_Greaterthan45);
raceAfAm = categorical(data1.race_African_American);
raceCauc = categorical(data1.race_Caucasian);
```

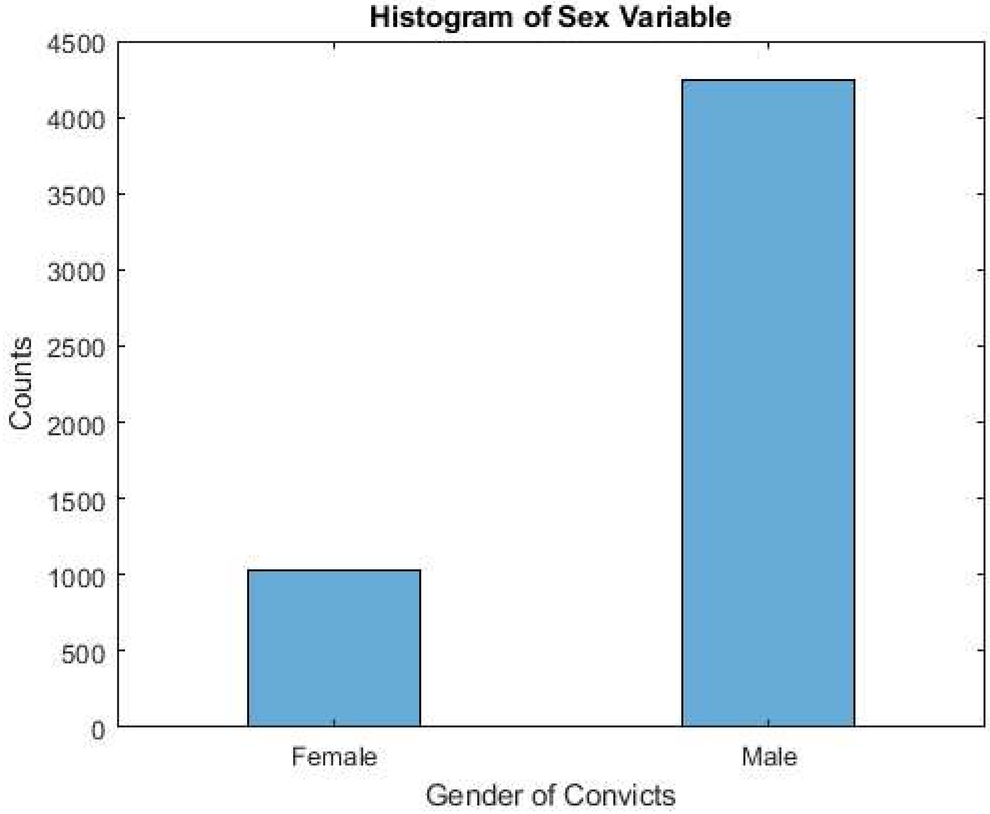
Create logical columns of of categorical variables for which the output % is equal to 1. %

```
female_recid = categorical(sex == '0' & targetcol == '1');
male_recid = categorical(sex == '1' & targetcol == '1');
agebelow25_recid = categorical(agebelow25 == '1' & targetcol == '1');
age25to45_recid =  categorical(age25to45 == '1' & targetcol == '1');
ageabove45_recid = categorical(ageabove45 == '1' & targetcol == '1');
raceAfAm_recid = categorical(raceAfAm == '1' & targetcol == '1');
raceCauc_recid = categorical(raceCauc == '1' & targetcol == '1');
```



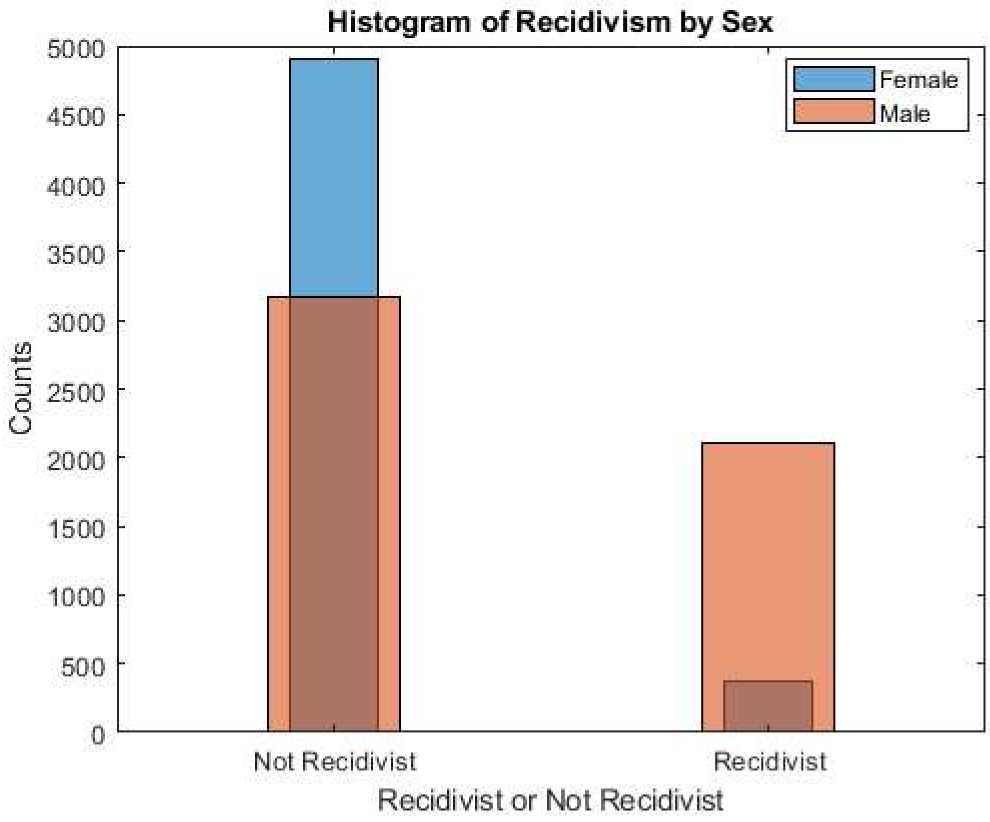
Plot histogram of the variable to explore the ratio of its values. %

```
figure
histogram(sex,'BarWidth',0.4);
title('Histogram of Sex Variable')
xlabel('Gender of Convicts')
ylabel('Counts')
xticklabels({'Female','Male'})
```



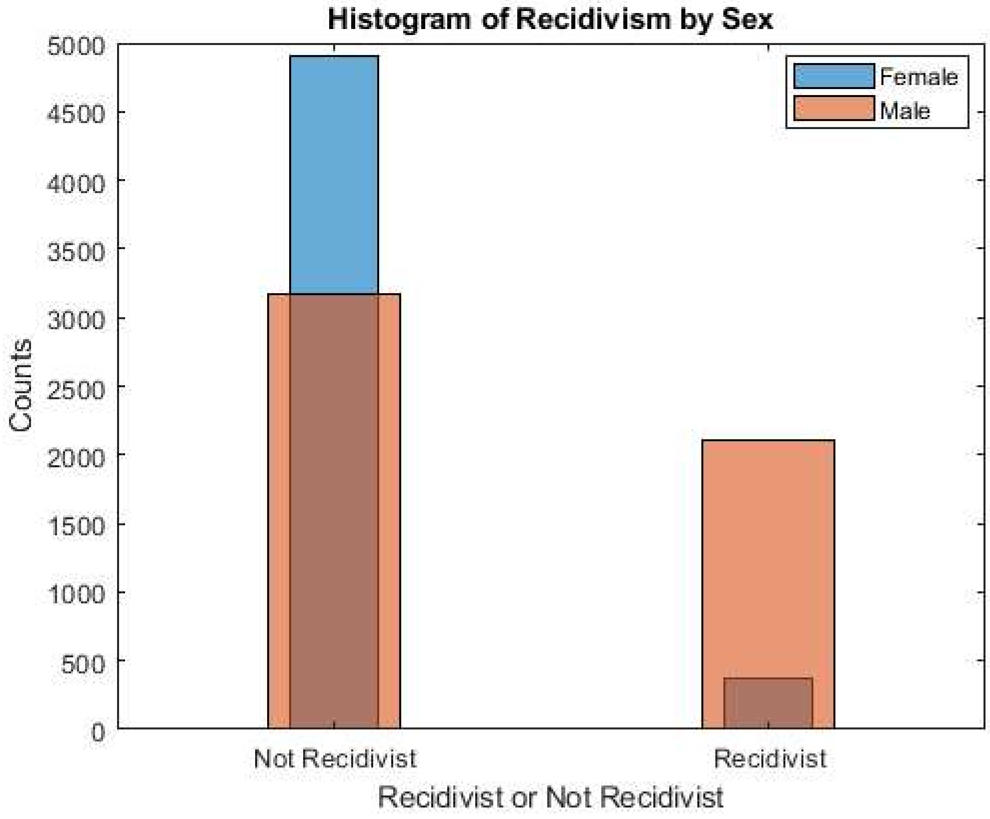
Plot histogram to compare the outputs for different values of a variable.

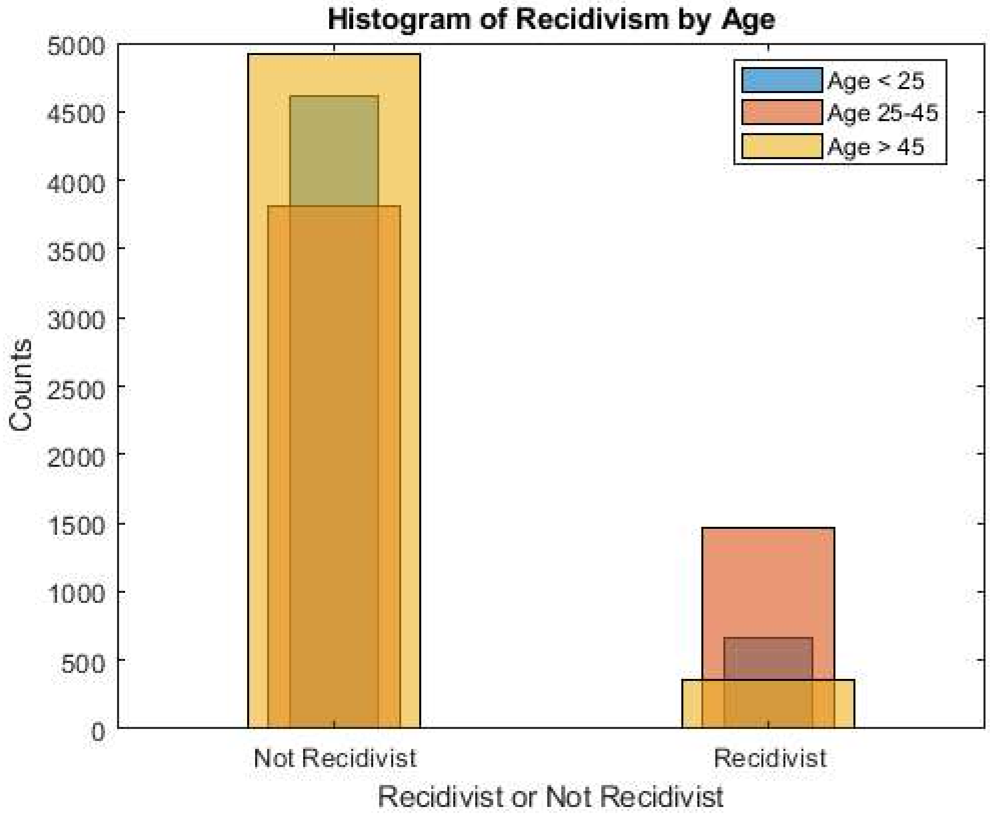
```
figure
histogram(female_recid,'BarWidth',0.2);
xticklabels({'Not Recidivist','Recidivist'})
hold on
histogram(male_recid,'BarWidth',0.3);
legend('Female', 'Male','Location','best')
title('Histogram of Recidivism by Sex')
xlabel('Recidivist or Not Recidivist')
ylabel('Counts')
hold off
```



Plot histogram to compare the outputs for different values of a variable.

```
figure
histogram(agebelow25_recid,'BarWidth',0.2);
xticklabels({'Not Recidivist','Recidivist'})
hold on
histogram(age25to45_recid,'BarWidth',0.3);
histogram(ageabove45_recid,'BarWidth',0.4);
legend('Age < 25', 'Age 25-45', 'Age > 45','Location','best')
title('Histogram of Recidivism by Age')
xlabel('Recidivist or Not Recidivist')
ylabel('Counts')
hold off
```

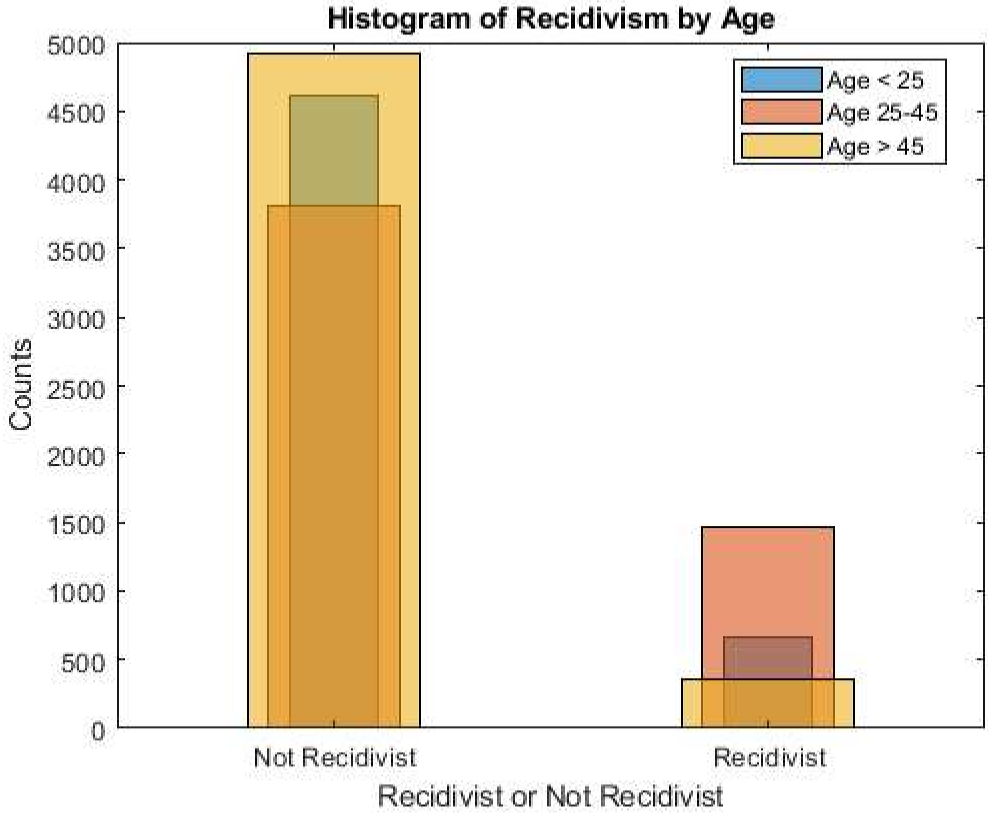


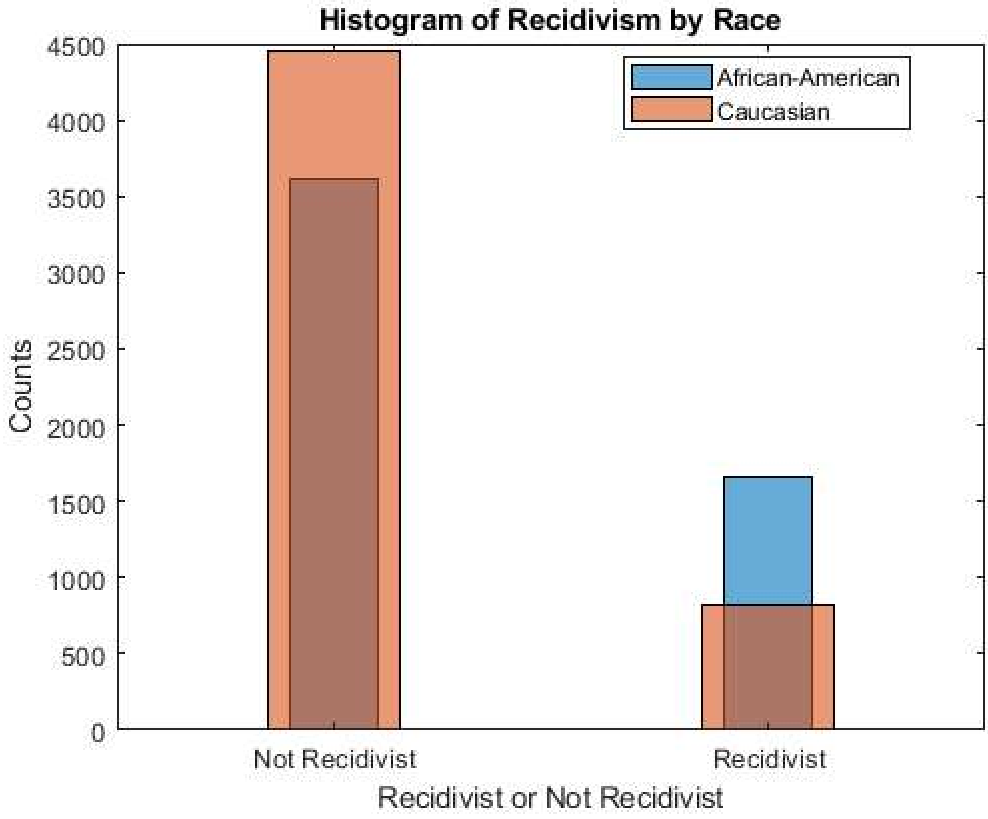


Plot histogram to compare the outputs for different values of a variable.

```
figure
histogram(raceAfAm_recid,'BarWidth',0.2);
xticklabels({'Not Recidivist','Recidivist'})
hold on
histogram(raceCauc_recid,'BarWidth',0.3);
legend('African-American', 'Caucasian','Location','best')
title('Histogram of Recidivism by Race')
```

```
xlabel('Recidivist or Not Recidivist')
ylabel('Counts')
hold off
```

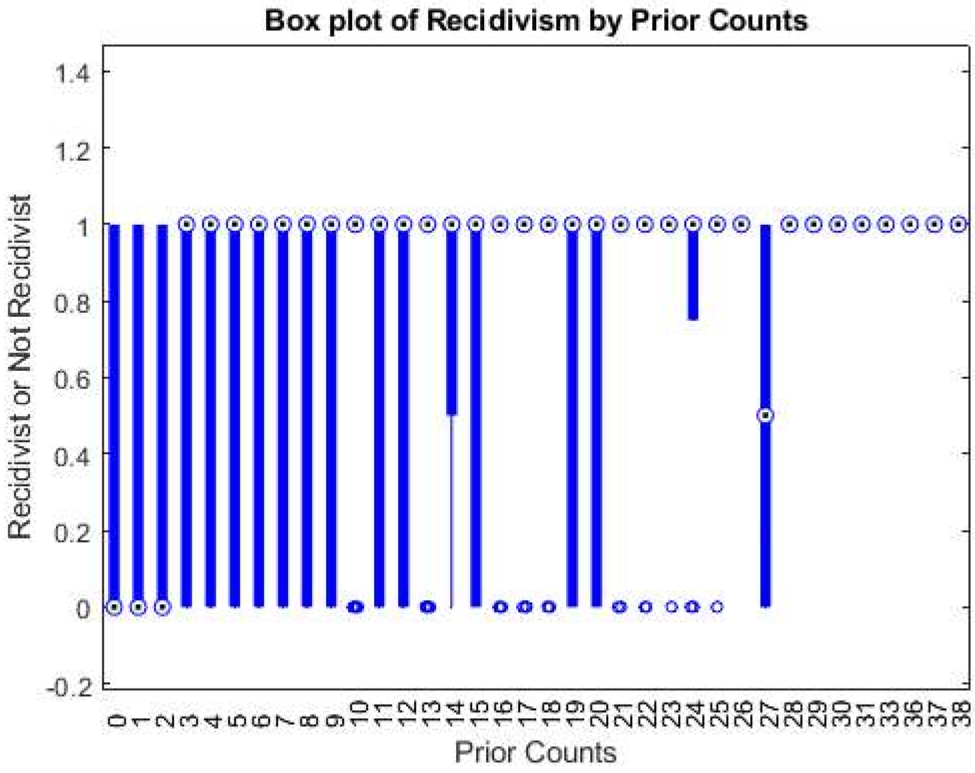




Plot boxplot to understand the relation of a continues variable with the binary output. %

```
figure
boxplot(data1.two_year_recid,data1.priors_count,'BoxStyle','filled','PlotStyle','compact')
```

```
title('Box plot of Recidivism by Prior Counts')
xlabel('Prior Counts')
ylabel('Recidivist or Not Recidivist')
```



Apart from the Exporatory Data Analysis done above, descriptive % statistics and a heatmap of correlation among variables would give % further clarity. But these can be done better in python than in MATLAB. %

```
% Export the data for further Exporatory Data Analysis in Python. %
% The code below is commented because it need not be executed repeatedly, %
% and can be executed by the Faculty assessing this script during the %
% assesment. %

% writetable(data1,'compas_for_EDA.csv');
```

After performing the remaining Exploratory data analysis in python, it was felt that there was no need to clean the data further. %

```
% Partition data into training and test sets using cross validation. %

% The code below is commented because it need not be executed repeatedly, %
% and can be executed by the Faculty assessing this script during the %
% assesment. %

%%rng(1);
%%cval1 = cvpartition(data1.two_year_recid,'HoldOut',0.2)
%%train_data1 = data1(training(cval1),:);
%%test_data1 = data1(test(cval1),:);
```

Export the partitioned training and test sets of data for modelling. %

```
% The code below is commented because it need not be executed repeatedly, %
% and can be executed by the Faculty assessing this script during the %
% assesment. %

%%writetable(train_data1,'Recidivismtrainset.csv');
%%writetable(test_data1,'Recidivismtestset.csv');
```

END %%

Published with MATLAB® R2021a