Open in Colab

Group307_kyphosis

1. Name - Arindam Dey
2. Group - 307 Group Members - ARINDAM DEY - 2020fc04251@wilp.bits-pilani.ac.in KAUSHIK DUBEY - 2020fc04245@wilp.bits-pilani.ac.in MOHAMMAD ATTAULLAH - 2020FC04274@wilp.bits-pilani.ac.in

3. Dataset Name - kyphosis.csv https://github.com/arindamdeyofficial/vertebraeSurgeryRiskFactor/blob/main/kyphosis.csv

RAW-
https://raw.githubusercontent.com/arindamdeyofficial/vertebraeSurgeryRiskFactor/7e5ba322ac93a5636528e286894521d35f7d4995/kyphosis.c

1. Import Libraries/Dataset

b. Import the required libraries

In [1]:
```
import pandas as pd
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

In [2]:
```
df = pd.read_csv('https://raw.githubusercontent.com/arindamdeyofficial/vertebraeSurgeryRiskFactor/7e5ba322ac93a
```

1. Data Visualization and Exploration

a. Print at least 5 rows for sanity check to identify all the features present in the dataset and if the target matches with them.

In [3]:
```
df.head()
```

Out[3]:

	Kyphosis	Age	Number	Start
0	absent	71	3	5
1	absent	158	3	14
2	present	128	4	5
3	absent	2	5	1
4	absent	1	4	15

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 81 entries, 0 to 80
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Kyphosis  81 non-null     object
 1   Age       81 non-null     int64
 2   Number    81 non-null     int64
 3   Start     81 non-null     int64
dtypes: int64(3), object(1)
memory usage: 2.7+ KB
```

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

Out[5]:
```
Index(['Kyphosis', 'Age', 'Number', 'Start'], dtype='object')
```

1. Data Visualization and Exploration

b. Print the description and shape of the dataset.

In [6]:
```
df.describe()
```

Out[6]:

	Age	Number	Start
count	81.000000	81.000000	81.000000
mean	83.654321	4.049383	11.493827
std	58.104251	1.619423	4.883962
min	1.000000	2.000000	1.000000
25%	26.000000	3.000000	9.000000
50%	87.000000	4.000000	13.000000
75%	130.000000	5.000000	16.000000
max	206.000000	10.000000	18.000000

In [7]:
```
df.shape
```

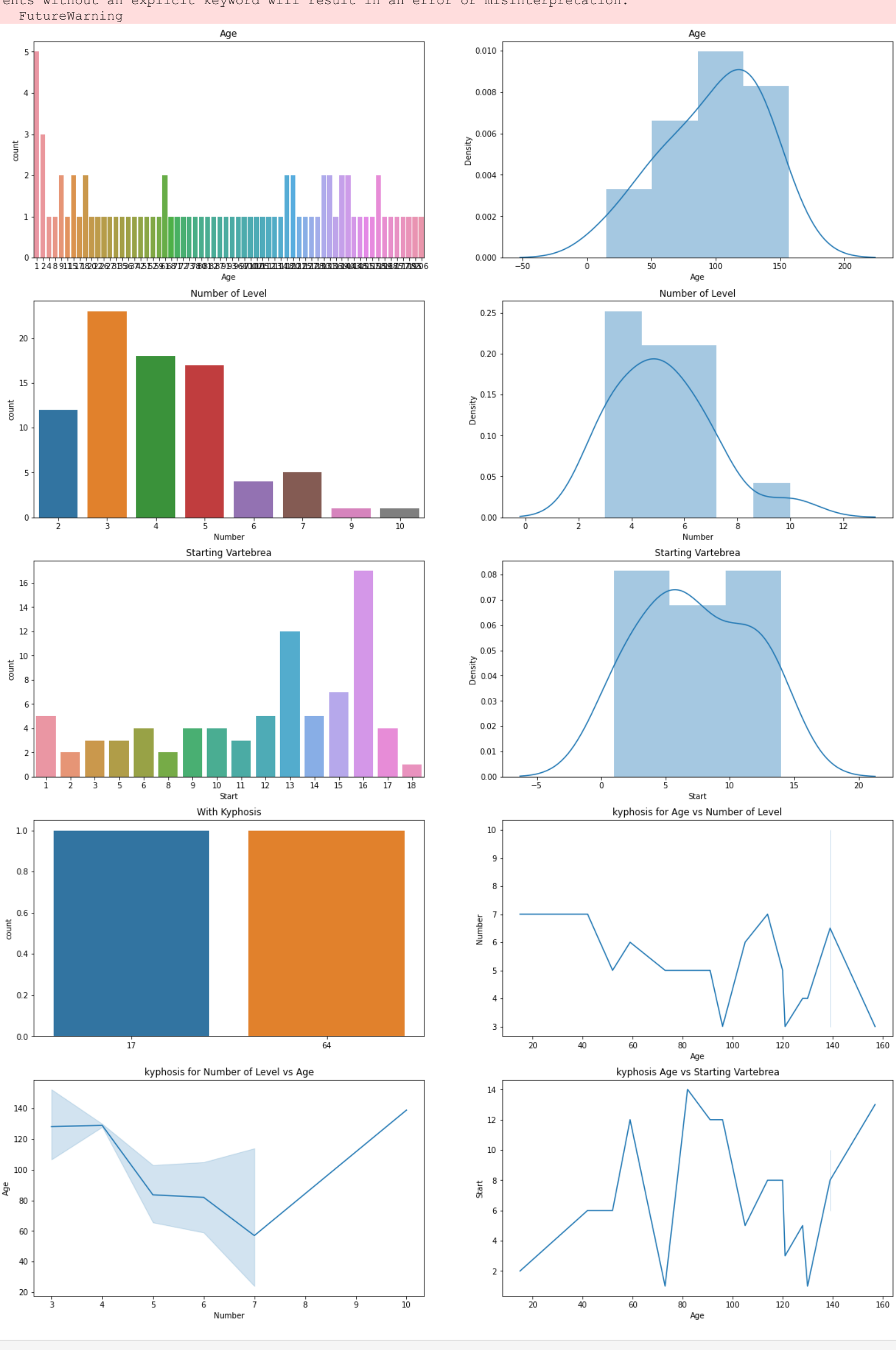
Out[7]:
```
(81, 4)
```

1. Data Visualization and Exploration

c. Provide appropriate visualization to get an insight about the dataset.

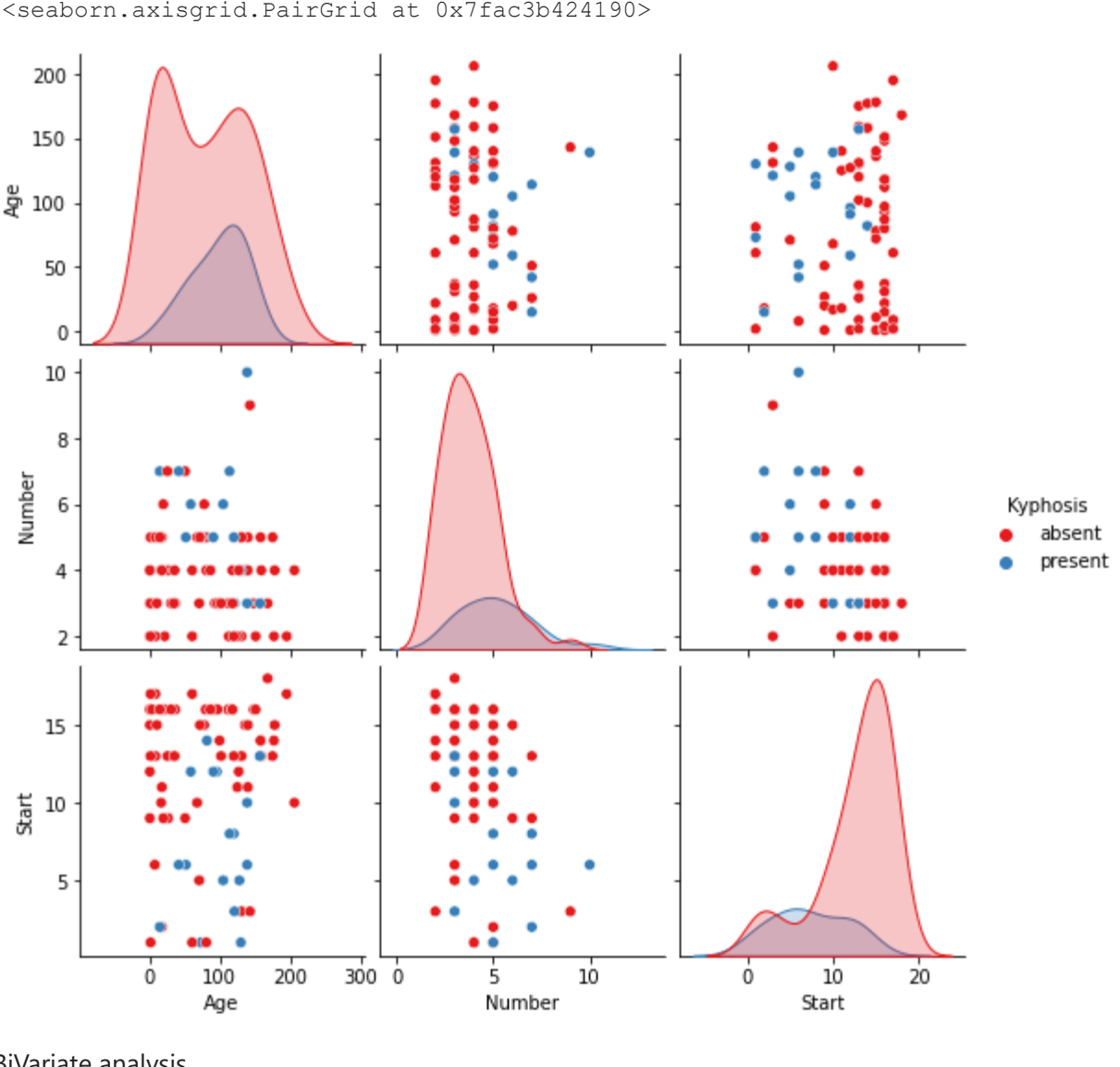
In [8]:
```
#Number === number of levels involved
#Start === starting vertebrae level of the surgery

n_rows=5
n_cols=2
width=20
height=30

fig,ax=plt.subplots(n_rows,n_cols,sharex=False,sharey=False,figsize=(width,height))
#Data normalisation view
wk = df[df["Kyphosis"]=="present"]

#Age
g = sns.countplot(df["Age"], ax=ax[0,0])
ax[0,0].set_title("Age")

g = sns.distplot(wk["Age"], ax=ax[0,1])
ax[0,1].set_title("Age")

#Number
g = sns.countplot(df["Number"], ax=ax[1,0])
ax[1,0].set_title("Number of Level")

g = sns.distplot(wk["Number"], ax=ax[1,1])
ax[1,1].set_title("Number of Level")

#Start
g = sns.countplot(df["Start"], ax=ax[2,0])
ax[2,0].set_title("Starting Vartebrea")

g = sns.distplot(wk["Start"], ax=ax[2,1])
ax[2,1].set_title("Starting Vartebrea")

#count kyphosis
l = len(df[df["Kyphosis"]=="present"])
g = sns.countplot([l,len(df)-l], ax=ax[3,0])
ax[3,0].set_title("With Kyphosis")

#Count kyphosis for Age vs Number of Level
#Having kyphosis Age vs Number Lineplot Age in X axis
wk = df[df["Kyphosis"]=="present"]
g = sns.lineplot(data=wk, x="Age", y="Number", ax=ax[3,1])
ax[3,1].set_title("kyphosis for Age vs Number of Level")

#Having kyphosis Number vs Age Countplot Number in X axis
g = sns.lineplot(data=wk, x="Number", y="Age", ax=ax[4,0])
ax[4,0].set_title("kyphosis for Number of Level vs Age")

#Having kyphosis Age vs Starting Vartebrea
g = sns.lineplot(data=wk, x="Age", y="Start",  ax=ax[4,1])
ax[4,1].set_title("kyphosis Age vs Starting Vartebrea")

plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2557: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2557: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2557: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
```



In [9]:
```
sns.pairplot(df,hue='Kyphosis',palette='Set1')
```

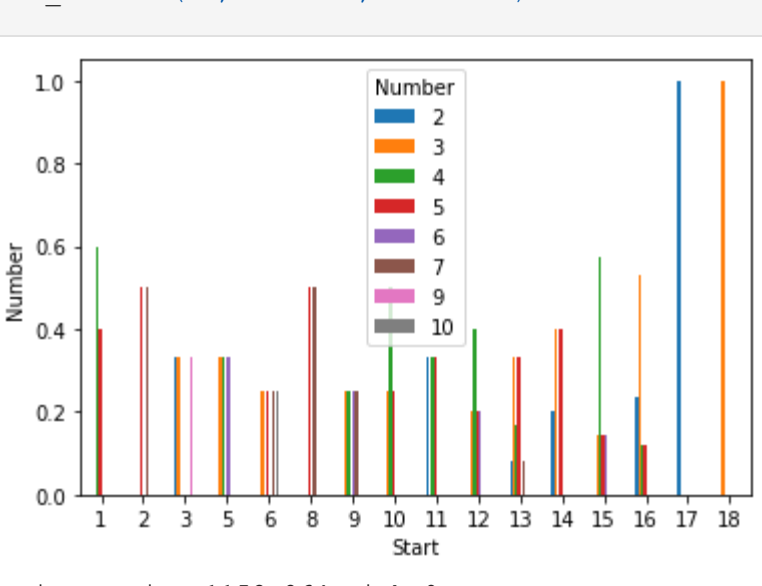
Out[9]:
```
<seaborn.axisgrid.PairGrid at 0x7fac3b424190>
```



BiVariate analysis

In [10]:
```
def bi_variate(df, col1, col2):
    cross_name=pd.crosstab(df[col1], df[col2], normalize='index')
    cross_name.plot.bar()
    plt.xlabel('{}'.format(col1))
    plt.ylabel('{}'.format(col2))
    plt.xticks(rotation=0)
    plt.figure(figsize=(16,12))
    plt.show()
    return cross_name*100
```

In [ ]:
```
bi_variate(df, 'Start', 'Number')
```




```
<Figure size 1152x864 with 0 Axes>
```

Out[ ]:

Number	2	3	4	5	6	7	9	10
Start								
1	0.000000	0.000000	60.000000	40.000000	0.000000	0.000000	0.000000	0.0
2	0.000000	0.000000	0.000000	50.000000	0.000000	50.000000	0.000000	0.0
3	33.333333	33.333333	0.000000	0.000000	0.000000	0.000000	33.333333	0.0
5	0.000000	33.333333	33.333333	0.000000	33.333333	0.000000	0.000000	0.0
6	0.000000	25.000000	0.000000	25.000000	0.000000	25.000000	0.000000	25.0
8	0.000000	0.000000	0.000000	50.000000	0.000000	50.000000	0.000000	0.0
9	0.000000	25.000000	25.000000	0.000000	25.000000	25.000000	0.000000	0.0
10	0.000000	25.000000	50.000000	25.000000	0.000000	0.000000	0.000000	0.0
11	33.333333	0.000000	33.333333	33.333333	0.000000	0.000000	0.000000	0.0
12	0.000000	20.000000	40.000000	20.000000	20.000000	0.000000	0.000000	0.0
13	8.333333	33.333333	16.666667	33.333333	0.000000	8.333333	0.000000	0.0
14	20.000000	40.000000	0.000000	40.000000	0.000000	0.000000	0.000000	0.0
15	0.000000	14.285714	57.142857	14.285714	14.285714	0.000000	0.000000	0.0
16	23.529412	52.941176	11.764706	11.764706	0.000000	0.000000	0.000000	0.0
17	100.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
18	0.000000	100.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0

d. Try exploring the data and see what insights can be drawn from the dataset.

Exploratory data analysis

1. Data Pre-processing and cleaning

a. Do the appropriate preprocessing of the data like identifying NULL or Missing Values if any, handling of outliers if present in the dataset, skewed data etc. Apply appropriate feature engineering techniques for them.

Now First check should be any Null is present or not

In [11]:
```
df.isna().any()
```

Out[11]:
```
Kyphosis    False
Age         False
Number      False
Start       False
dtype: bool
```

In [12]:
```
df.isna().sum()
```

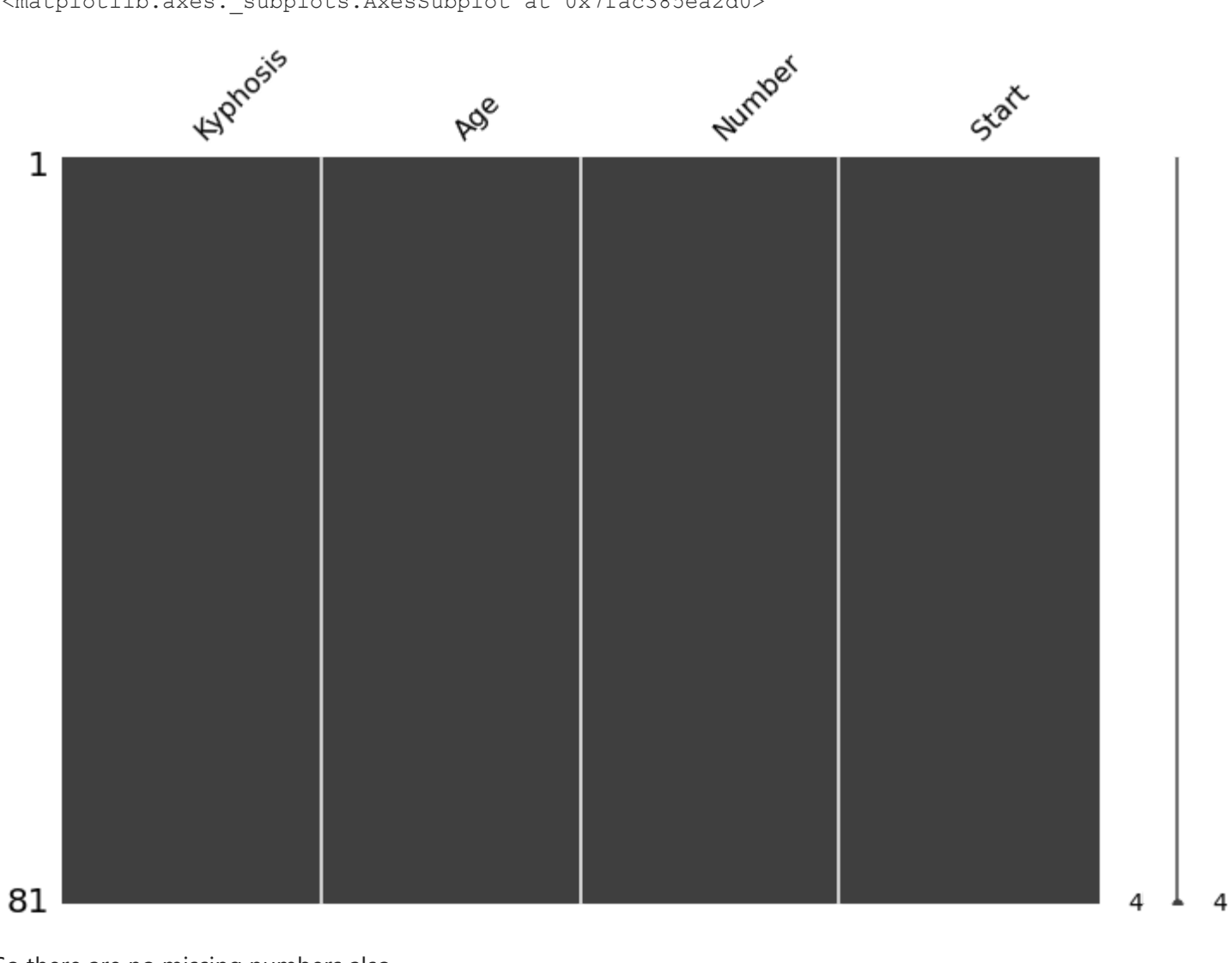
Out[12]:
```
Kyphosis    0
Age         0
Number      0
Start       0
dtype: int64
```

Now there are no Null values present

In [13]:
```
!pip install missingno
import missingno
missingno.matrix(df,figsize=(12,8))
```

```
Requirement already satisfied: missingno in /usr/local/lib/python3.7/dist-packages (0.5.0)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.7/dist-packages (from missingno) (3.2.2)
Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from missingno) (1.4.1)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from missingno) (1.19.5)
Requirement already satisfied: seaborn in /usr/local/lib/python3.7/dist-packages (from missingno) (0.11.1)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packages (from matplotlib->missingno) (0.10.0)
Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib->missingno) (2.8.2)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib->missingno) (2.4.7)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib->missingno) (1.3.1)
Requirement already satisfied: six in /usr/local/lib/python3.7/dist-packages (from cycler>=0.10->matplotlib->missingno) (1.15.0)
Requirement already satisfied: pandas>=0.23 in /usr/local/lib/python3.7/dist-packages (from seaborn->missingno) (1.1.5)
Requirement already satisfied: pytz>=2017.2 in /usr/local/lib/python3.7/dist-packages (from pandas>=0.23->seaborn->missingno) (2018.9)
```

Out[13]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fac385ea2d0>
```



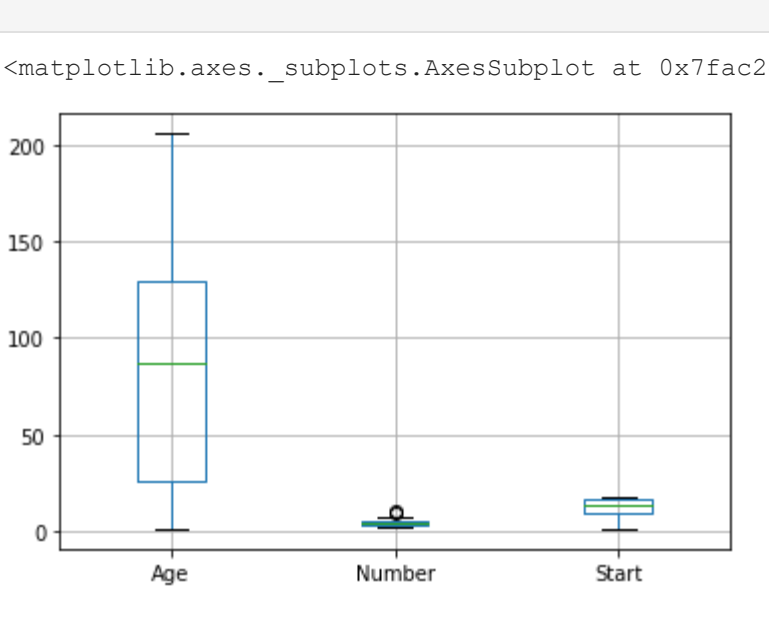
So there are no missing numbers also

Next Step: Outlier Identificayion and removal

Box plot use the Inter Quantile Range(IQR) method. This is easier so will be using

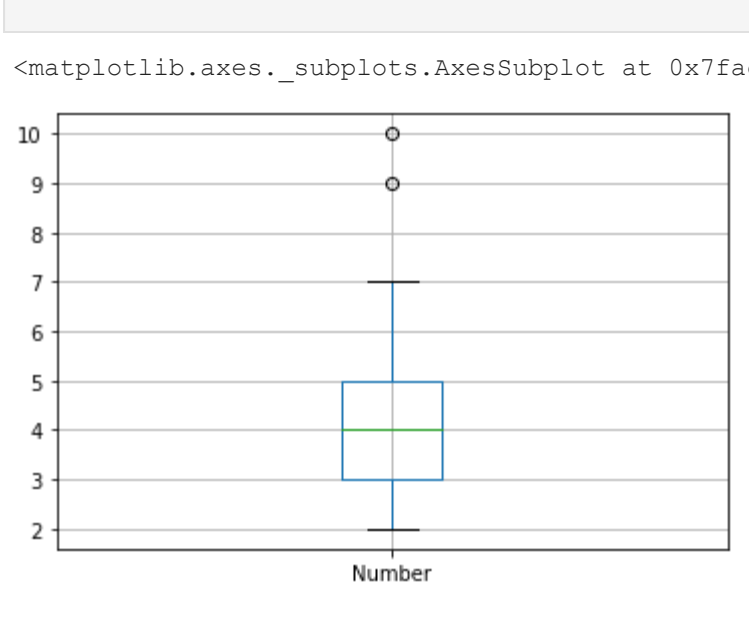
In [14]:
```
df.boxplot()
```

Out[14]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fac2e1e5490>
```



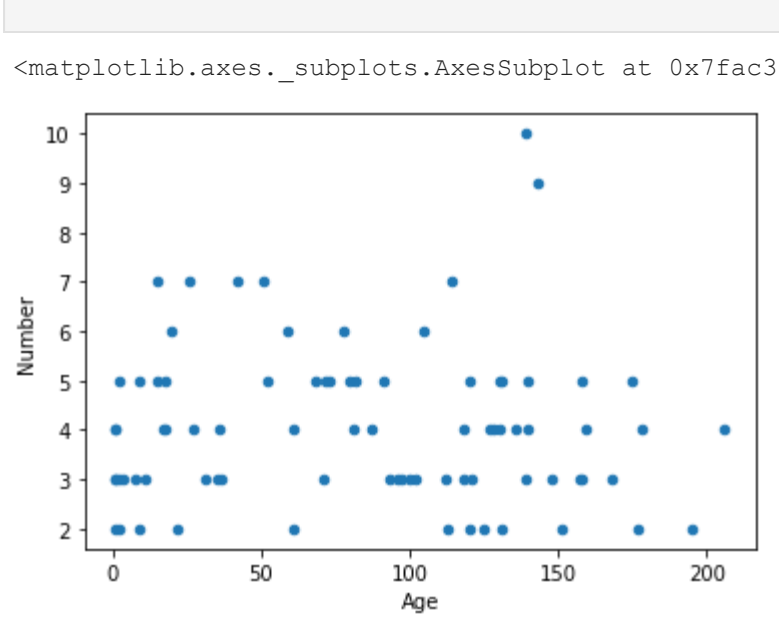
In [15]:
```
df.boxplot(column=['Number'])
```

Out[15]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fac2da4ec50>
```



In [16]:
```
df.plot.scatter('Age', 'Number')
```

Out[16]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fac385e19d0>
```



Z-score is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured Through thi method also we can identify Outlier

In [17]:
```
mean = np.mean(df["Number"])
mean
```

Out[17]:
```
4.049382716049383
```

In [18]:
```
std = np.std(df["Number"])
std
```

Out[18]:
```
1.6093955200472556
```

In [19]:
```
threshold = 2
outlierCombo = []
outlier = []
zScore = []
for i in df["Number"]:
    z = (i-mean)/std
    if z > threshold:
      outlierCombo.append([i,z])
      outlier.append(i)
      zScore.append(z)
outlierCombo
```

Out[19]:
```
[[9, 3.076072489505411], [10, 3.697423790378075]]
```

In [20]:
```
df[df["Number"]>=9]
```

Out[20]:

	Kyphosis	Age	Number	Start
42	absent	143	9	3
52	present	139	10	6

This 2 are really unrealistic Age people live according to age also and as per scatter plot and ZScore. But data is skewed toward 100++ years of age and data quantity also less so prefferable not drop but replace in 3 sigma method the 2 outliers.

In [21]:
```
leftVal = mean - 3*std
rightVal = mean + 3*std
print(mean, std, leftVal,rightVal)
outlL = df[df["Number"]<leftVal]
outlR = df[df["Number"]> rightVal]
print(outlL["Number"].count(), outlR["Number"].count())
print("threshold value count: ", df[df["Number"] == rightVal]["Number"].count())
print(outlR)
```

```
4.049382716049383 1.6093955200472556 -0.7788038440923835 8.87756927619115
0 2
threshold value count:  0
   Kyphosis  Age  Number  Start
42   absent  143       9      3
52  present  139      10      6
```

In [22]:
```
df.loc[df["Number"] > rightVal, "Number"] = rightVal
```

In [23]:
```
df[df["Number"] > rightVal]
```

Out[23]:

Kyphosis	Age	Number	Start

In [24]:
```
print("threshold value count: ", df[df["Number"] == rightVal]["Number"].count())
```

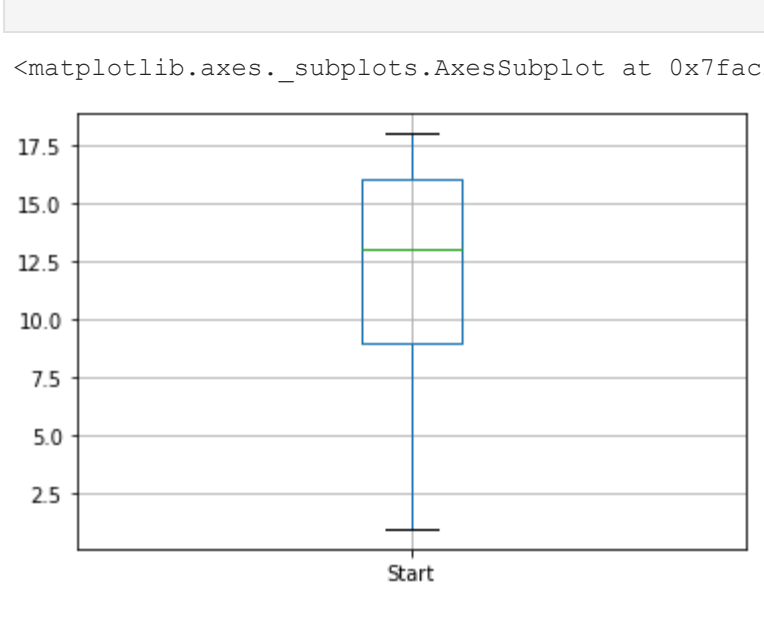
```
threshold value count:  2
```

In [25]:
```
index_names = df.loc[df["Number"]>=9].index
index_names
```

Out[25]:
```
Int64Index([], dtype='int64')
```

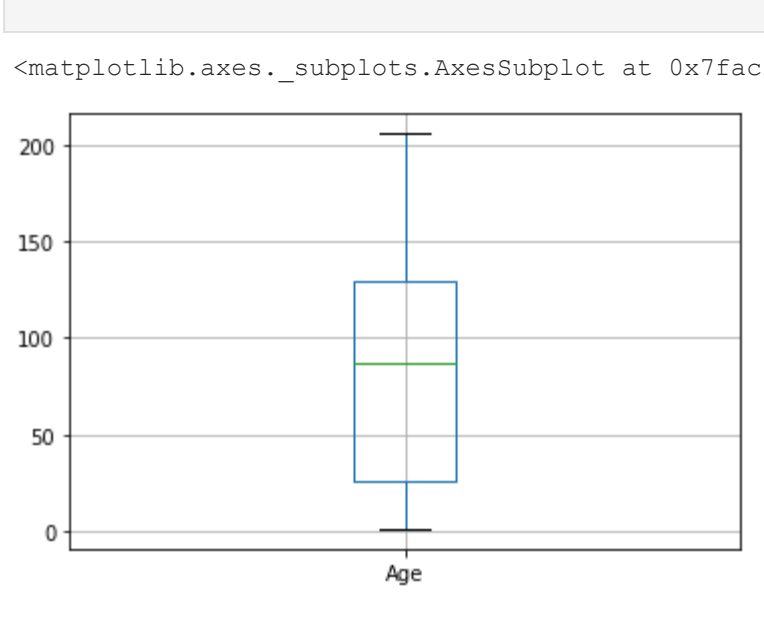
In [26]:
```
df.boxplot(column=['Start'])
```

Out[26]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fac2d93f190>
```



In [27]:
```
df.boxplot(column=['Age'])
```

Out[27]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fac2d8cd550>
```



Now we need to see Group by Age how other data is behaving

And population distribution distributes values in ranges according to data density. It is needed as Age data is right skewed

In [28]:
```
ageSeries = df.copy(deep=True)
ageSeries["AgeRanges"] = pd.qcut(df["Age"], 10)
ageSeries
```

Out[28]:

	Kyphosis	Age	Number	Start	AgeRanges
0	absent	71	3.0	5	(68.0, 87.0]
1	absent	158	3.0	14	(139.0, 158.0]
2	present	128	4.0	5	(125.0, 139.0]
3	absent	2	5.0	1	(0.999, 4.0]
4	absent	1	4.0	15	(0.999, 4.0]
...	...	...	...	...	...
76	present	157	3.0	13	(139.0, 158.0]
77	absent	26	7.0	13	(18.0, 36.0]
78	absent	120	2.0	13	(112.0, 125.0]
79	present	42	7.0	6	(36.0, 68.0]
80	absent	36	4.0	13	(18.0, 36.0]

81 rows × 5 columns

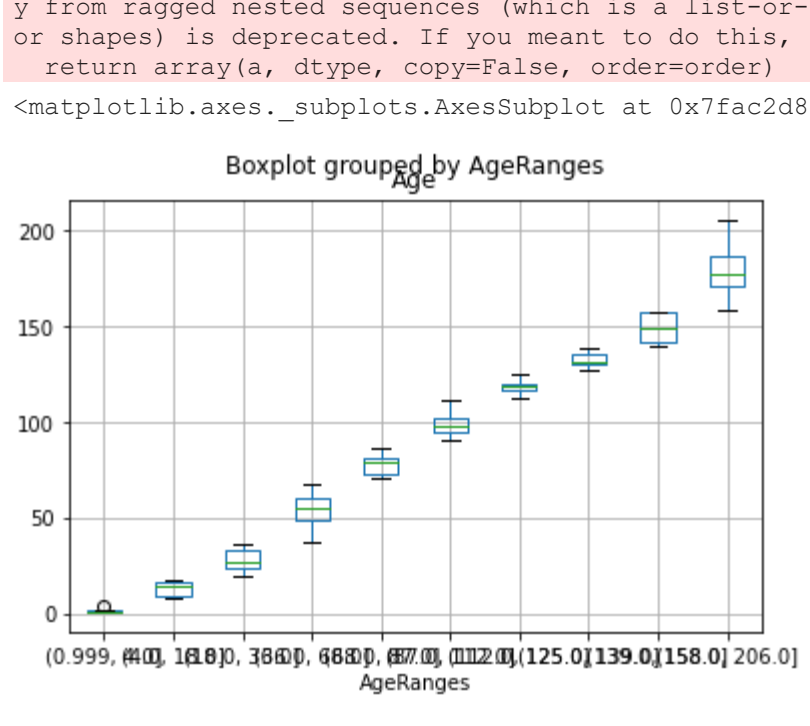
In [29]:
```
age = df["Age"]
age.max()
```

Out[29]:
```
206
```

In [30]:
```
ageSeries.boxplot(column=['Age'], by="AgeRanges")
```

```
/usr/local/lib/python3.7/dist-packages/numpy/core/_asarray.py:83: VisibleDeprecationWarning: Creating an ndarray from ragged nested sequences (which is a list-or-tuple of lists-or-tuples-or ndarrays with different lengths or shapes) is deprecated. If you meant to do this, you must specify 'dtype=object' when creating the ndarray
  return array(a, dtype, copy=False, order=order)
```

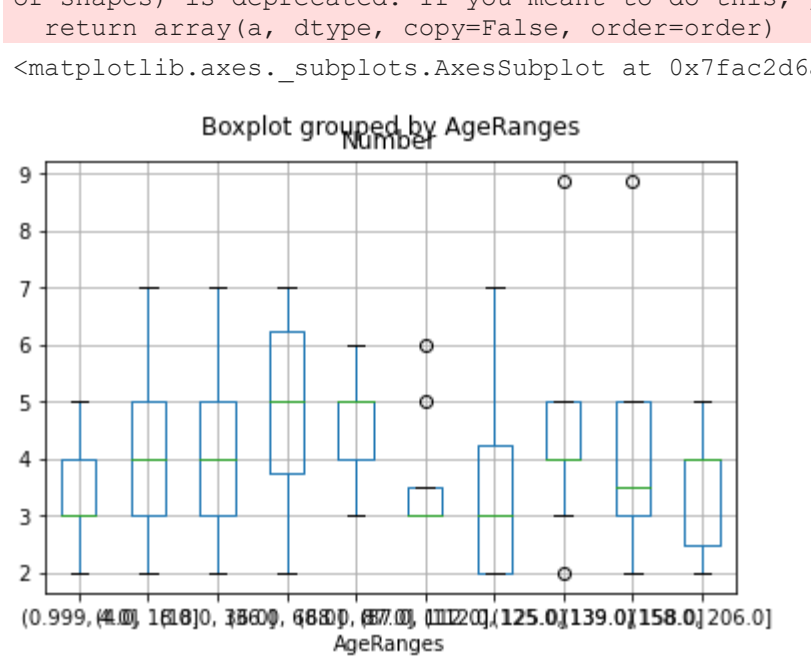
Out[30]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fac2d831a10>
```



In [31]:
```
ageSeries.boxplot(column=['Number'], by="AgeRanges")
```

```
/usr/local/lib/python3.7/dist-packages/numpy/core/_asarray.py:83: VisibleDeprecationWarning: Creating an ndarray from ragged nested sequences (which is a list-or-tuple of lists-or-tuples-or ndarrays with different lengths or shapes) is deprecated. If you meant to do this, you must specify 'dtype=object' when creating the ndarray
  return array(a, dtype, copy=False, order=order)
```

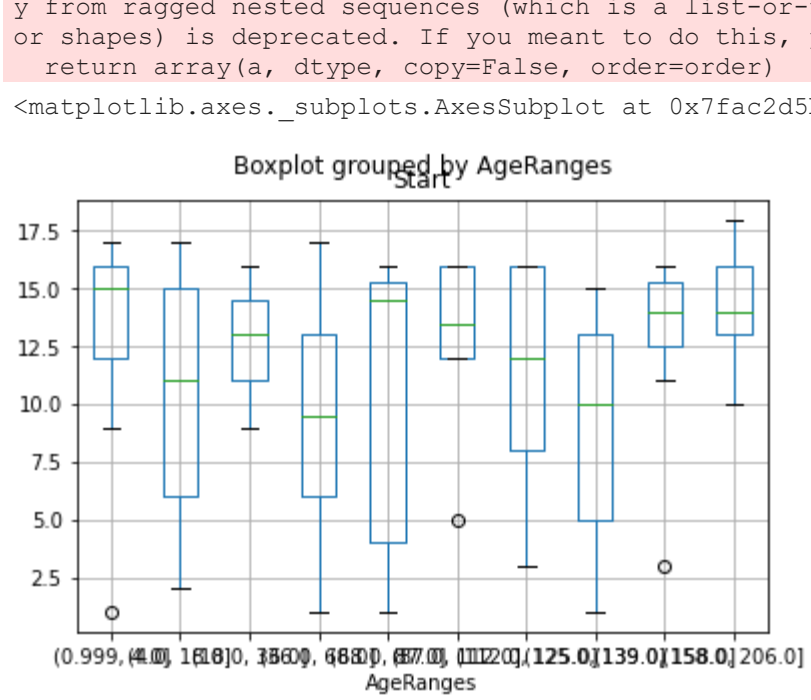
Out[31]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fac2d6ac4d0>
```



In [32]:
```
ageSeries.boxplot(column=['Start'], by="AgeRanges")
```

```
/usr/local/lib/python3.7/dist-packages/numpy/core/_asarray.py:83: VisibleDeprecationWarning: Creating an ndarray from ragged nested sequences (which is a list-or-tuple of lists-or-tuples-or ndarrays with different lengths or shapes) is deprecated. If you meant to do this, you must specify 'dtype=object' when creating the ndarray
  return array(a, dtype, copy=False, order=order)
```

Out[32]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fac2d5b9510>
```



b. Apply the feature transformation techniques like Standardization, Normalization, etc. You are free to apply the appropriate transformations depending upon the structure and the complexity of your dataset.

Replace categorical Column with Numeric value

Column segregartion

In [33]:
```
df.head(10)
#Encoding (categorical to Numerical)
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder

#target column
target_col=['Kyphosis']

#categorical cols
categorical_cols=df.nunique()[df.nunique() < 5].keys().tolist()
categorical_cols=[x for x in categorical_cols if x not in target_col]
print("categorical columns: {}".format(categorical_cols))

#numerical cols
cont_cols=[x for x in df.columns if x not in categorical_cols+target_col]
print("Numerical columns: {}".format(cont_cols))

#binary cols
bin_cols=df.nunique()[df.nunique() == 2].keys().tolist()
print("Binary columns: {}".format(bin_cols))

#multi cols(more than 2 values)
mul_cols=[x for x in categorical_cols if x not in bin_cols]
print("Multi valued columns: {}".format(mul_cols))
```

```
categorical columns: []
Numerical columns: ['Age', 'Number', 'Start']
Binary columns: ['Kyphosis']
Multi valued columns: []
```

In [34]:
```
#Label Encoding for Binary cols
le=LabelEncoder()
for i in bin_cols:
    df[i]=le.fit_transform(df[i])

#Duplicating cols for multi-value cols
#df=pd.get_dummies(data=df,columns=mul_cols)
df.head(10)
```

Out[34]:

	Kyphosis	Age	Number	Start
0	0	71	3.0	5
1	0	158	3.0	14
2	1	128	4.0	5
3	0	2	5.0	1
4	0	1	4.0	15
5	0	1	2.0	16
6	0	61	2.0	17
7	0	37	3.0	16
8	0	113	2.0	16
9	1	59	6.0	12

1. Data Pre-processing and cleaning

c. Do the correlational analysis on the dataset. Provide a visualization for the same.

Feature Selection

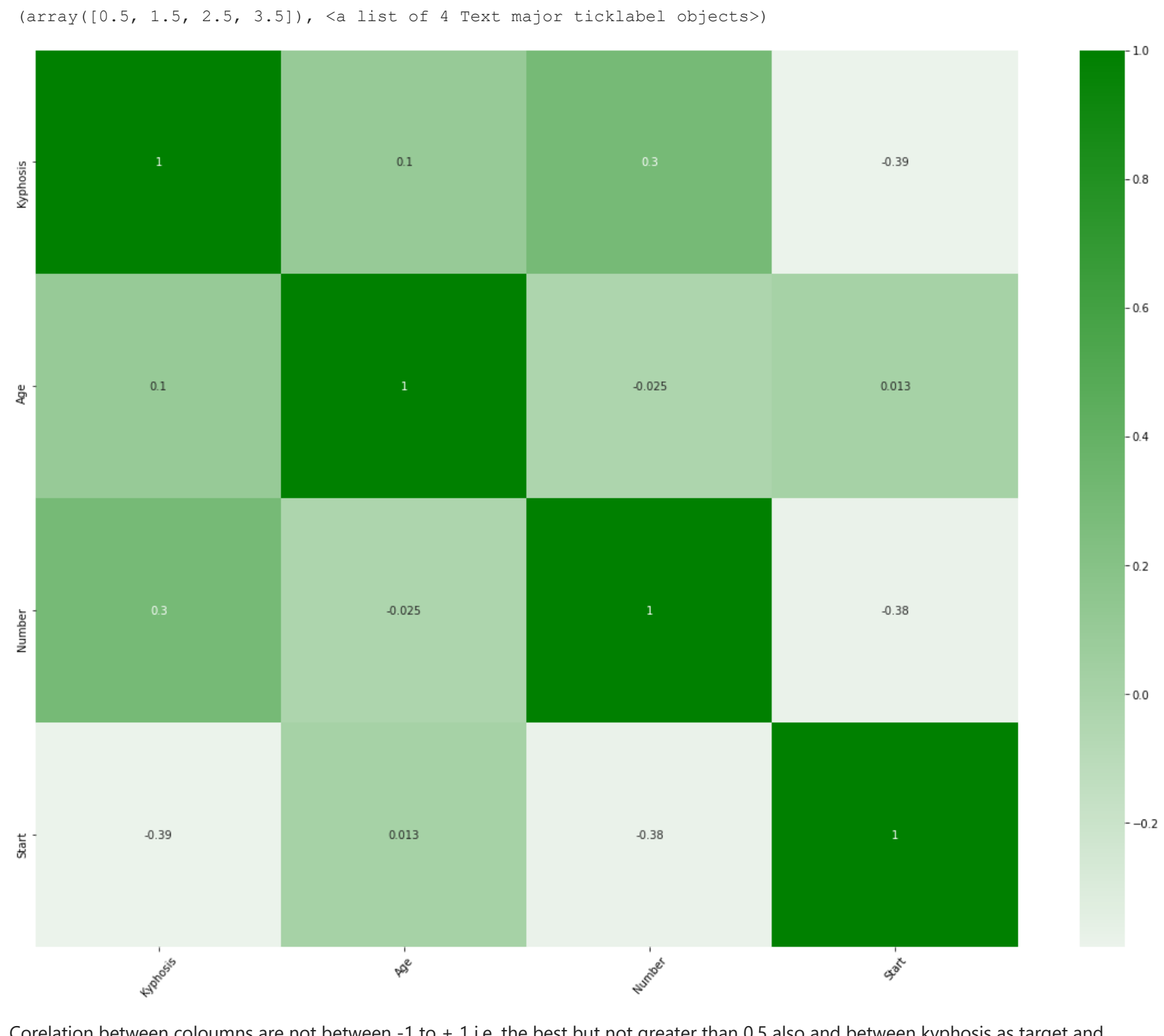
We can do by corelation. mostly continuous data as input and categorical data as output. SO we can use below methods.

1. ANOVA correlation coefficient (linear)
2. Kendall's rank coefficient (nonlinear)

kendall

In [35]:
```
#Co-relation
#correlation analysis shows us how to determine both the nature and strength of relationship between two varial
#correaltion lies between -1 to 1 (0: No corelation; -1: perfect negative correlation; +1: positive core)
correlation=df.corr(method='kendall')
correlation

#plotting correlation Matrix
plt.figure(figsize=(20,15))
cm = sns.light_palette("green", as_cmap=True)
sns.heatmap(correlation, xticklabels=correlation.columns.values, yticklabels=correlation.columns.values, cmap=c
plt.xticks(rotation=50)
```

Out[35]: (array([0.5, 1.5, 2.5, 3.5]), <a list of 4 Text major ticklabel objects>)



Corelation between coloumns are not between -1 to + 1 i.e. the best but not greater than 0.5 also and between kyphosis as target and other column is generally greater. so we are good.

ANOVA F measure

In [38]:
```
import scipy.stats as stats
# stats f_oneway functions takes the groups as input and returns ANOVA F and p value
fvalue, pvalue = stats.f_oneway(df['Age'], df['Number'], df['Start'])
print(fvalue, pvalue)
```

```
138.10440440436454 1.2198309671987674e-40
```

p value is e to the power 40 i.e. very small. So null hypothesis rejected. Variance of these columns are not same. So they are less corelated. We can take all 3 to build model.

1. Data Preparation

a. Do the final feature selection and extract them into Column X and the class label into Column into Y.

In [39]:
```
x=df.drop('Kyphosis',axis=1)
y=df['Kyphosis']
```

1. Data Preparation

b. Split the dataset into training and test sets.

In [40]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.33,random_state=101)
```

Part B

1. Model Building

a. Perform Model Development using at least three models, separately. You are free to apply any Machine Learning Models on the dataset. Deep Learning Models are strictly not allowed.

We will try Descision Tree as data is too small and number of column is also less preferrebly. And also try Random forest and Logistics regression as all of them are Classification models.

We have general common classification techniques

1. Logistic Regression
2. Naïve Bayes
3. Stochastic Gradient Descent
4. K-Nearest Neighbours
5. Decision Tree
6. Random Forest
7. Support Vector Machine

1. Model Building

Logistics Regression

In [41]:
```
from sklearn.metrics import classification_report,confusion_matrix
def viewConfusionMatrix(y_test, y_pred):
  print(confusion_matrix(y_test,y_pred))
  print(classification_report(y_test,y_pred))
```

In [42]:
```
from sklearn.linear_model import LogisticRegression

def logicalRegres(x,y,x_train,y_train,x_test,y_test):
  clf = LogisticRegression(random_state=0).fit(x, y)
  logreg = LogisticRegression()
  logreg.fit(x_train, y_train)
  y_pred = logreg.predict(x_test)
  print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(logreg.score(x_test, y_test)))
  return y_pred
```

Now logical regression using all column

In [43]:
```
y_pred = logicalRegres(x,y,x_train,y_train,x_test,y_test)
```

```
Accuracy of logistic regression classifier on test set: 0.74
```

In [44]:
```
viewConfusionMatrix(y_test, y_pred)
```

```
[[19  0]
 [ 7  1]]
              precision    recall  f1-score   support

           0       0.73      1.00      0.84        19
           1       1.00      0.12      0.22         8

    accuracy                           0.74        27
   macro avg       0.87      0.56      0.53        27
weighted avg       0.81      0.74      0.66        27
```

Logical regression not using "Start" column

Need to prepare other sets of x i.e. input as Start corelation is not very good and will compare them in same Logical Regression model. Their accuracy and performance.

In [45]:
```
x_no_start=df.drop(['Kyphosis', 'Start'],axis=1)
```

In [46]:
```
from sklearn.model_selection import train_test_split
x_no_start_train,x_no_start_test,y_no_start_train,y_no_start_test=train_test_split(x_no_start,y,test_size=0.33,
```

In [47]:
```
y_no_start_pred = logicalRegres(x_no_start,y,x_no_start_train,y_no_start_train,x_no_start_test,y_no_start_test)
```

```
Accuracy of logistic regression classifier on test set: 0.74
```

In [48]:
```
viewConfusionMatrix(y_no_start_test, y_no_start_pred)
```

```
[[19  0]
 [ 7  1]]
              precision    recall  f1-score   support

           0       0.73      1.00      0.84        19
           1       1.00      0.12      0.22         8

    accuracy                           0.74        27
   macro avg       0.87      0.56      0.53        27
weighted avg       0.81      0.74      0.66        27
```

So no change in accuracy though. SO we are good

DecisionTree

In [49]:
```
from sklearn.tree import DecisionTreeClassifier
dtree=DecisionTreeClassifier()
#b.      Train the model and print the training accuracy and loss values
dtree.fit(x_train,y_train)
#2.     Performance Evaluation
#b. Do the prediction for the test data and display the results for the inference
y_pred=dtree.predict(x_test)
#a. Print the confusion matrix. Provide appropriate analysis for the same.
```

a. Print the confusion matrix. Provide appropriate analysis for the same.

In [50]:
```
viewConfusionMatrix(y_test, y_pred)
```

```
[[15  4]
 [ 6  2]]
              precision    recall  f1-score   support

           0       0.71      0.79      0.75        19
           1       0.33      0.25      0.29         8

    accuracy                           0.63        27
   macro avg       0.52      0.52      0.52        27
weighted avg       0.60      0.63      0.61        27
```

RandomForest

In [51]:
```
from sklearn.ensemble import RandomForestClassifier
rfc=RandomForestClassifier(n_estimators=20000)
rfc.fit(x_train,y_train)
y_pred=rfc.predict(x_test)
```

In [52]:
```
viewConfusionMatrix(y_test, y_pred)
```

```
[[19  0]
 [ 6  2]]
              precision    recall  f1-score   support

           0       0.76      1.00      0.86        19
           1       1.00      0.25      0.40         8

    accuracy                           0.78        27
   macro avg       0.88      0.62      0.63        27
weighted avg       0.83      0.78      0.73        27
```

data is not very much imbalanced though but we can try SMOTE approach

In [53]:
```
!pip install imbalanced-learn
import imblearn
print(imblearn.__version__)
from collections import Counter
from imblearn.over_sampling import SMOTE
from imblearn.under_sampling import RandomUnderSampler
from imblearn.pipeline import Pipeline
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RepeatedStratifiedKFold
```

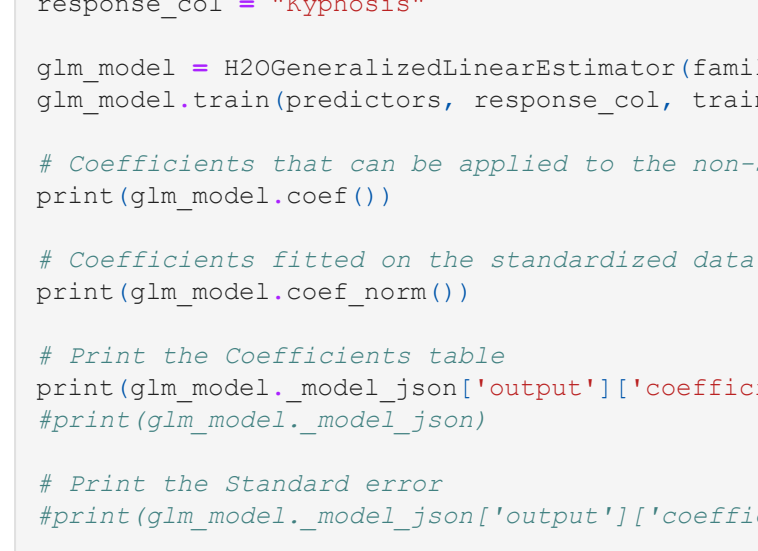
```
Requirement already satisfied: imbalanced-learn in /usr/local/lib/python3.7/dist-packages (0.4.3)
Requirement already satisfied: scikit-learn>=0.20 in /usr/local/lib/python3.7/dist-packages (from imbalanced-le
arn) (0.22.2.post1)
Requirement already satisfied: scipy>=0.13.3 in /usr/local/lib/python3.7/dist-packages (from imbalanced-learn)
(1.4.1)
Requirement already satisfied: numpy>=1.8.2 in /usr/local/lib/python3.7/dist-packages (from imbalanced-learn)
(1.19.5)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/dist-packages (from scikit-learn>=0.20-
>imbalanced-learn) (1.0.1)
0.4.3
/usr/local/lib/python3.7/dist-packages/sklearn/externals/six.py:31: FutureWarning: The module is deprecated in
version 0.21 and will be removed in version 0.23 since we've dropped support for Python 2.7. Please rely on the
official version of six (https://pypi.org/project/six/).
  "(https://pypi.org/project/six/).", FutureWarning)
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:144: FutureWarning: The sklearn.neighbors.b
ase module is  deprecated in version 0.22 and will be removed in version 0.24. The corresponding classes / func
tions should instead be imported from sklearn.neighbors. Anything that cannot be imported from sklearn.neighbor
s is now part of the private API.
  warnings.warn(message, FutureWarning)
```

In [54]:
```
x=df.drop(['Kyphosis'],axis=1)
y=df['Kyphosis']
```

In [55]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.33,random_state=101)
```

In [56]:
```
oversample = SMOTE()
counter = Counter(y)
print(counter)
X, y = oversample.fit_resample(x, y)
# summarize the new class distribution
counter = Counter(y)
print(counter)
# scatter plot of examples by class label
for label, _ in counter.items():
        row_ix = np.where(y == label)[0]
        plt.scatter(X[row_ix, 0], X[row_ix, 1], label=str(label))
plt.legend()
plt.show()
```

```
Counter({0: 64, 1: 17})
Counter({0: 64, 1: 64})
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function safe_indexing i
s deprecated; safe_indexing is deprecated in version 0.22 and will be removed in version 0.24.
  warnings.warn(msg, category=FutureWarning)
```



Generalized Linear Model Table will help for linear regression and predict drop column. It is needed as between target i.e. kyphosis and one of the X i.e column "Start" corelation is not so good. And same between column "Start" and "Number" .

Generalized Linear Models (GLM) estimate regression models for outcomes following exponential distributions. In addition to the Gaussian (i.e. normal) distribution, these include Poisson, binomial, and gamma distributions. Each serves a different purpose, and depending on distribution and link function choice, can be used either for prediction or classification.

In [37]:
```
from google.colab import files
df.to_csv(r"GlmData.csv", sep=',', index=False)
files.download('GlmData.csv')
```

In [59]:
```
!pip install h2o
import h2o
h2o.init()
from h2o.estimators.glm import H2OGeneralizedLinearEstimator
glm = h2o.import_file("https://raw.githubusercontent.com/arindamdeyofficial/vertebraeSurgeryRiskFactor/main/Glm
print(glm.as_data_frame())
glm["Kyphosis"] = glm["Age"].asfactor()
glm["Age"] = glm["Age"].asfactor()
glm["Number"] = glm["Age"].asfactor()
glm["Start"] = glm["Age"].asfactor()
predictors  = ["Age", "Number", "Start"]
response_col = "Kyphosis"

glm_model = H2OGeneralizedLinearEstimator(family= "AUTO")
glm_model.train(predictors, response_col, training_frame= glm)

# Coefficients that can be applied to the non-standardized data.
print(glm_model.coef())

# Coefficients fitted on the standardized data (requires standardize = True, which is on by default)
print(glm_model.coef_norm())

# Print the Coefficients table
print(glm_model._model_json['output']['coefficients_table'])
#print(glm_model._model_json)

# Print the Standard error
#print(glm_model._model_json['output']['coefficients_table']['std_error'])

# Print the p values
#print(glm_model._model_json['output']['coefficients_table']['p_value'])

# Print the z values
#print(glm_model._model_json['output']['coefficients_table']['z_value'])

# Retrieve a graphical plot of the standardized coefficient magnitudes
#glm_model.std_coef_plot()
```

```
Requirement already satisfied: h2o in /usr/local/lib/python3.7/dist-packages (3.32.1.6)
Requirement already satisfied: colorama>=0.3.8 in /usr/local/lib/python3.7/dist-packages (from h2o) (0.4.4)
Requirement already satisfied: tabulate in /usr/local/lib/python3.7/dist-packages (from h2o) (0.8.9)
Requirement already satisfied: future in /usr/local/lib/python3.7/dist-packages (from h2o) (0.16.0)
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (from h2o) (2.23.0)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-packages (from requests->h2o) (1.24.3)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests->h2o) (2.10)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests->h2o) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests->h2o) (2021.5.30)
Checking whether there is an H2O instance running at http://localhost:54321 . connected.
```

H2O_cluster_uptime:	1 min 48 secs
H2O_cluster_timezone:	Etc/UTC
H2O_data_parsing_timezone:	UTC
H2O_cluster_version:	3.32.1.6
H2O_cluster_version_age:	8 days
H2O_cluster_name:	H2O_from_python_unknownUser_6b6kg0
H2O_cluster_total_nodes:	1
H2O_cluster_free_memory:	3.171 Gb
H2O_cluster_total_cores:	2
H2O_cluster_allowed_cores:	2
H2O_cluster_status:	locked, healthy
H2O_connection_url:	http://localhost:54321
H2O_connection_proxy:	{"http": null, "https": null}
H2O_internal_security:	False
H2O_API_Extensions:	Amazon S3, XGBoost, Algos, AutoML, Core V3, TargetEncoder, Core V4
Python_version:	3.7.11 final

```
Parse progress: |████████████████████████████████████████████| 100%
    Kyphosis  Age  Number  Start
0          0   71     3.0      5
1          0  158     3.0     14
2          1  128     4.0      5
3          0    2     5.0      1
4          0    1     4.0     15
..       ...  ...     ...    ...
76         1  157     3.0     13
77         0   26     7.0     13
78         0  120     2.0     13
79         1   42     7.0      6
80         0   36     4.0     13

[81 rows x 4 columns]
glm Model Build progress: |███████████████████████████████████████| 100%
```

{"coefs_class_1": {"Intercept": -3.6545389744338364, "Age.1": 1.6390739405780042, "Age.2": 0.0, "Age.4": 0.0, "Age.8": 0.0, "Age.9": 0.0, "Age.11": 0.0, "Age.15": 0.0, "Age.17": 0.0, "Age.18": 0.0, "Age.20": 0.0, "Age.22": 0.0, "Age.26": 0.0, "Age.27": 0.0, "Age.31": 0.0, "Age.35": 0.0, "Age.36": 0.0, "Age.37": 0.0, "Age.42": 0.0, ...

t.36': 0.0, 'Start.37': 0.0, 'Start.42': 0.0, 'Start.51': 0.0, 'Start.52': 0.0, 'Start.59': 0.0, 'Start.61': 0.
0, 'Start.68': 0.0, 'Start.71': 0.0, 'Start.72': 0.0, 'Start.73': 0.0, 'Start.78': 0.0, 'Start.80': 0.0, 'Star
t.81': 0.0, 'Start.82': 0.0, 'Start.87': 0.0, 'Start.91': 0.0, 'Start.93': 0.0, 'Start.96': 0.0, 'Start.97': 0.
0, 'Start.100': 0.0, 'Start.102': 0.0, 'Start.105': 0.0, 'Start.112': 0.0, 'Start.113': 0.0, 'Start.114': 0.0,
 'Start.118': 0.0, 'Start.120': 0.0, 'Start.121': 0.0, 'Start.125': 0.0, 'Start.127': 0.0, 'Start.128': 0.0, 'S
tart.130': 0.0, 'Start.131': 0.0, 'Start.136': 0.0, 'Start.139': 0.0, 'Start.140': 0.0, 'Start.143': 0.0, 'Star
t.148': 0.0, 'Start.151': 0.0, 'Start.157': 0.0, 'Start.158': 0.0, 'Start.159': 0.0, 'Start.168': 0.0, 'Start.1
75': 0.8124179954006394, 'Start.177': 0.0, 'Start.178': 0.0, 'Start.195': 0.0, 'Start.206': 0.0}, 'std_coefs_cl
ass_177': {'Intercept': -4.037332382285116, 'Age.1': 0.0, 'Age.2': 0.0, 'Age.4': 0.0, 'Age.8': 0.0, 'Age.9': 0.
0, 'Age.11': 0.0, 'Age.15': 0.0, 'Age.17': 0.0, 'Age.18': 0.0, 'Age.20': 0.0, 'Age.22': 0.0, 'Age.26': 0.0, 'Ag
e.27': 0.0, 'Age.31': 0.0, 'Age.35': 0.0, 'Age.36': 0.0, 'Age.37': 0.0, 'Age.42': 0.0, 'Age.51': 0.0, 'Age.52':
 0.0, 'Age.59': 0.0, 'Age.61': 0.0, 'Age.68': 0.0, 'Age.71': 0.0, 'Age.72': 0.0, 'Age.73': 0.0, 'Age.78': 0.0,
 'Age.80': 0.0, 'Age.81': 0.0, 'Age.82': 0.0, 'Age.87': 0.0, 'Age.91': 0.0, 'Age.93': 0.0, 'Age.96': 0.0, 'Age.
97': 0.0, 'Age.100': 0.0, 'Age.102': 0.0, 'Age.105': 0.0, 'Age.112': 0.0, 'Age.113': 0.0, 'Age.114': 0.0, 'Age.
118': 0.0, 'Age.120': 0.0, 'Age.121': 0.0, 'Age.125': 0.0, 'Age.127': 0.0, 'Age.128': 0.0, 'Age.130': 0.0, 'Ag
e.131': 0.0, 'Age.136': 0.0, 'Age.139': 0.0, 'Age.140': 0.0, 'Age.143': 0.0, 'Age.148': 0.0, 'Age.151': 0.0, 'A
ge.157': 0.0, 'Age.158': 0.0, 'Age.159': 0.0, 'Age.168': 0.0, 'Age.175': 0.0, 'Age.177': 0.8124183414923933, 'A
ge.178': 0.0, 'Age.195': 0.0, 'Age.206': 0.0, 'Number.1': 0.0, 'Number.2': 0.0, 'Number.4': 0.0, 'Number.8': 0.
0, 'Number.9': 0.0, 'Number.11': 0.0, 'Number.15': 0.0, 'Number.17': 0.0, 'Number.18': 0.0, 'Number.20': 0.0,
 'Number.22': 0.0, 'Number.26': 0.0, 'Number.27': 0.0, 'Number.31': 0.0, 'Number.35': 0.0, 'Number.36': 0.0, 'N
umber.37': 0.0, 'Number.42': 0.0, 'Number.51': 0.0, 'Number.52': 0.0, 'Number.59': 0.0, 'Number.61': 0.0, 'Numb
er.68': 0.0, 'Number.71': 0.0, 'Number.72': 0.0, 'Number.73': 0.0, 'Number.78': 0.0, 'Number.80': 0.0, 'Number.
81': 0.0, 'Number.82': 0.0, 'Number.87': 0.0, 'Number.91': 0.0, 'Number.93': 0.0, 'Number.96': 0.0, 'Number.9
7': 0.0, 'Number.100': 0.0, 'Number.102': 0.0, 'Number.105': 0.0, 'Number.112': 0.0, 'Number.113': 0.0, 'Numbe
r.114': 0.0, 'Number.118': 0.0, 'Number.120': 0.0, 'Number.121': 0.0, 'Number.125': 0.0, 'Number.127': 0.0, 'Nu
mber.128': 0.0, 'Number.130': 0.0, 'Number.131': 0.0, 'Number.136': 0.0, 'Number.139': 0.0, 'Number.140': 0.0,
 'Number.143': 0.0, 'Number.148': 0.0, 'Number.151': 0.0, 'Number.157': 0.0, 'Number.158': 0.0, 'Number.159':
 0.0, 'Number.168': 0.0, 'Number.175': 0.0, 'Number.177': 0.8124183414923922, 'Number.178': 0.0, 'Number.195':
 0.0, 'Number.206': 0.0, 'Start.1': 0.0, 'Start.2': 0.0, 'Start.4': 0.0, 'Start.8': 0.0, 'Start.9': 0.0, 'Star
t.11': 0.0, 'Start.15': 0.0, 'Start.17': 0.0, 'Start.18': 0.0, 'Start.20': 0.0, 'Start.22': 0.0, 'Start.26': 0.
0, 'Start.27': 0.0, 'Start.31': 0.0, 'Start.35': 0.0, 'Start.36': 0.0, 'Start.37': 0.0, 'Start.42': 0.0, 'Star
t.51': 0.0, 'Start.52': 0.0, 'Start.59': 0.0, 'Start.61': 0.0, 'Start.68': 0.0, 'Start.71': 0.0, 'Start.72': 0.
0, 'Start.73': 0.0, 'Start.78': 0.0, 'Start.80': 0.0, 'Start.81': 0.0, 'Start.82': 0.0, 'Start.87': 0.0, 'Star
t.91': 0.0, 'Start.93': 0.0, 'Start.96': 0.0, 'Start.97': 0.0, 'Start.100': 0.0, 'Start.102': 0.0, 'Start.105':
 0.0, 'Start.112': 0.0, 'Start.113': 0.0, 'Start.114': 0.0, 'Start.118': 0.0, 'Start.120': 0.0, 'Start.121': 0.
0, 'Start.125': 0.0, 'Start.127': 0.0, 'Start.128': 0.0, 'Start.130': 0.0, 'Start.131': 0.0, 'Start.136': 0.0,
 'Start.139': 0.0, 'Start.140': 0.0, 'Start.143': 0.0, 'Start.148': 0.0, 'Start.151': 0.0, 'Start.157': 0.0, 'S
tart.158': 0.0, 'Start.159': 0.0, 'Start.168': 0.0, 'Start.175': 0.0, 'Start.177': 0.8124183414923936, 'Start.1
78': 0.0, 'Start.195': 0.0, 'Start.206': 0.0}, 'std_coefs_class_178': {'Intercept': -4.037332939540057, 'Age.
1': 0.0, 'Age.2': 0.0, 'Age.4': 0.0, 'Age.8': 0.0, 'Age.9': 0.0, 'Age.11': 0.0, 'Age.15': 0.0, 'Age.17': 0.0,
 'Age.18': 0.0, 'Age.20': 0.0, 'Age.22': 0.0, 'Age.26': 0.0, 'Age.27': 0.0, 'Age.31': 0.0, 'Age.35': 0.0, 'Age.
36': 0.0, 'Age.37': 0.0, 'Age.42': 0.0, 'Age.51': 0.0, 'Age.52': 0.0, 'Age.59': 0.0, 'Age.61': 0.0, 'Age.68':
 0.0, 'Age.71': 0.0, 'Age.72': 0.0, 'Age.73': 0.0, 'Age.78': 0.0, 'Age.80': 0.0, 'Age.81': 0.0, 'Age.82': 0.0,
 'Age.87': 0.0, 'Age.91': 0.0, 'Age.93': 0.0, 'Age.96': 0.0, 'Age.97': 0.0, 'Age.100': 0.0, 'Age.102': 0.0, 'Ag
e.105': 0.0, 'Age.112': 0.0, 'Age.113': 0.0, 'Age.114': 0.0, 'Age.118': 0.0, 'Age.120': 0.0, 'Age.121': 0.0, 'A
ge.125': 0.0, 'Age.127': 0.0, 'Age.128': 0.0, 'Age.130': 0.0, 'Age.131': 0.0, 'Age.136': 0.0, 'Age.139': 0.0,
 'Age.140': 0.0, 'Age.143': 0.0, 'Age.148': 0.0, 'Age.151': 0.0, 'Age.157': 0.0, 'Age.158': 0.0, 'Age.159': 0.
0, 'Age.168': 0.0, 'Age.175': 0.0, 'Age.177': 0.0, 'Age.178': 0.8124186492446404, 'Age.195': 0.0, 'Age.206': 0.
0, 'Number.1': 0.0, 'Number.2': 0.0, 'Number.4': 0.0, 'Number.8': 0.0, 'Number.9': 0.0, 'Number.11': 0.0, 'Numb
er.15': 0.0, 'Number.17': 0.0, 'Number.18': 0.0, 'Number.20': 0.0, 'Number.22': 0.0, 'Number.26': 0.0, 'Number.
27': 0.0, 'Number.31': 0.0, 'Number.35': 0.0, 'Number.36': 0.0, 'Number.37': 0.0, 'Number.42': 0.0, 'Number.5
1': 0.0, 'Number.52': 0.0, 'Number.59': 0.0, 'Number.61': 0.0, 'Number.68': 0.0, 'Number.71': 0.0, 'Number.72':
 0.0, 'Number.73': 0.0, 'Number.78': 0.0, 'Number.80': 0.0, 'Number.81': 0.0, 'Number.82': 0.0, 'Number.87': 0.
0, 'Number.91': 0.0, 'Number.93': 0.0, 'Number.96': 0.0, 'Number.97': 0.0, 'Number.100': 0.0, 'Number.102': 0.
0, 'Number.105': 0.0, 'Number.112': 0.0, 'Number.113': 0.0, 'Number.114': 0.0, 'Number.118': 0.0, 'Number.120':
 0.0, 'Number.121': 0.0, 'Number.125': 0.0, 'Number.127': 0.0, 'Number.128': 0.0, 'Number.130': 0.0, 'Number.13
1': 0.0, 'Number.136': 0.0, 'Number.139': 0.0, 'Number.140': 0.0, 'Number.143': 0.0, 'Number.148': 0.0, 'Numbe
r.151': 0.0, 'Number.157': 0.0, 'Number.158': 0.0, 'Number.159': 0.0, 'Number.168': 0.0, 'Number.175': 0.0, 'Nu
mber.177': 0.0, 'Number.178': 0.812418649244641, 'Number.195': 0.0, 'Number.206': 0.0, 'Start.1': 0.0, 'Start.
2': 0.0, 'Start.4': 0.0, 'Start.8': 0.0, 'Start.9': 0.0, 'Start.11': 0.0, 'Start.15': 0.0, 'Start.17': 0.0, 'St
art.18': 0.0, 'Start.20': 0.0, 'Start.22': 0.0, 'Start.26': 0.0, 'Start.27': 0.0, 'Start.31': 0.0, 'Start.35':
 0.0, 'Start.36': 0.0, 'Start.37': 0.0, 'Start.42': 0.0, 'Start.51': 0.0, 'Start.52': 0.0, 'Start.59': 0.0, 'St
art.61': 0.0, 'Start.68': 0.0, 'Start.71': 0.0, 'Start.72': 0.0, 'Start.73': 0.0, 'Start.78': 0.0, 'Start.80':
 0.0, 'Start.81': 0.0, 'Start.82': 0.0, 'Start.87': 0.0, 'Start.91': 0.0, 'Start.93': 0.0, 'Start.96': 0.0, 'St
art.97': 0.0, 'Start.100': 0.0, 'Start.102': 0.0, 'Start.105': 0.0, 'Start.112': 0.0, 'Start.113': 0.0, 'Start.
114': 0.0, 'Start.118': 0.0, 'Start.120': 0.0, 'Start.121': 0.0, 'Start.125': 0.0, 'Start.127': 0.0, 'Start.12
8': 0.0, 'Start.130': 0.0, 'Start.131': 0.0, 'Start.136': 0.0, 'Start.139': 0.0, 'Start.140': 0.0, 'Start.143':
 0.0, 'Start.148': 0.0, 'Start.151': 0.0, 'Start.157': 0.0, 'Start.158': 0.0, 'Start.159': 0.0, 'Start.168': 0.
0, 'Start.175': 0.0, 'Start.177': 0.0, 'Start.178': 0.8124186492446406, 'Start.195': 0.0, 'Start.206': 0.0}, 's
td_coefs_class_195': {'Intercept': -4.0373384371916625, 'Age.1': 0.0, 'Age.2': 0.0, 'Age.4': 0.0, 'Age.8': 0.0,
 'Age.9': 0.0, 'Age.11': 0.0, 'Age.15': 0.0, 'Age.17': 0.0, 'Age.18': 0.0, 'Age.20': 0.0, 'Age.22': 0.0, 'Age.2
6': 0.0, 'Age.27': 0.0, 'Age.31': 0.0, 'Age.35': 0.0, 'Age.36': 0.0, 'Age.37': 0.0, 'Age.42': 0.0, 'Age.51': 0.
0, 'Age.52': 0.0, 'Age.59': 0.0, 'Age.61': 0.0, 'Age.68': 0.0, 'Age.71': 0.0, 'Age.72': 0.0, 'Age.73': 0.0, 'Ag
e.78': 0.0, 'Age.80': 0.0, 'Age.81': 0.0, 'Age.82': 0.0, 'Age.87': 0.0, 'Age.91': 0.0, 'Age.93': 0.0, 'Age.96':
 0.0, 'Age.97': 0.0, 'Age.100': 0.0, 'Age.102': 0.0, 'Age.105': 0.0, 'Age.112': 0.0, 'Age.113': 0.0, 'Age.114':
 0.0, 'Age.118': 0.0, 'Age.120': 0.0, 'Age.121': 0.0, 'Age.125': 0.0, 'Age.127': 0.0, 'Age.128': 0.0, 'Age.13
0': 0.0, 'Age.131': 0.0, 'Age.136': 0.0, 'Age.139': 0.0, 'Age.140': 0.0, 'Age.143': 0.0, 'Age.148': 0.0, 'Age.1
51': 0.0, 'Age.157': 0.0, 'Age.158': 0.0, 'Age.159': 0.0, 'Age.168': 0.0, 'Age.175': 0.0, 'Age.177': 0.0, 'Age.
178': 0.0, 'Age.195': 0.8124189135387114, 'Age.206': 0.0, 'Number.1': 0.0, 'Number.2': 0.0, 'Number.4': 0.0, 'N
umber.8': 0.0, 'Number.9': 0.0, 'Number.11': 0.0, 'Number.15': 0.0, 'Number.17': 0.0, 'Number.18': 0.0, 'Numbe
r.20': 0.0, 'Number.22': 0.0, 'Number.26': 0.0, 'Number.27': 0.0, 'Number.31': 0.0, 'Number.35': 0.0, 'Number.3
6': 0.0, 'Number.37': 0.0, 'Number.42': 0.0, 'Number.51': 0.0, 'Number.52': 0.0, 'Number.59': 0.0, 'Number.61':
 0.0, 'Number.68': 0.0, 'Number.71': 0.0, 'Number.72': 0.0, 'Number.73': 0.0, 'Number.78': 0.0, 'Number.80': 0.
0, 'Number.81': 0.0, 'Number.82': 0.0, 'Number.87': 0.0, 'Number.91': 0.0, 'Number.93': 0.0, 'Number.96': 0.0,
 'Number.97': 0.0, 'Number.100': 0.0, 'Number.102': 0.0, 'Number.105': 0.0, 'Number.112': 0.0, 'Number.113': 0.
0, 'Number.114': 0.0, 'Number.118': 0.0, 'Number.120': 0.0, 'Number.121': 0.0, 'Number.125': 0.0, 'Number.127':
 0.0, 'Number.128': 0.0, 'Number.130': 0.0, 'Number.131': 0.0, 'Number.136': 0.0, 'Number.139': 0.0, 'Number.14
0': 0.0, 'Number.143': 0.0, 'Number.148': 0.0, 'Number.151': 0.0, 'Number.157': 0.0, 'Number.158': 0.0, 'Numbe
r.159': 0.0, 'Number.168': 0.0, 'Number.175': 0.0, 'Number.177': 0.0, 'Number.178': 0.0, 'Number.195': 0.812418
9135387123, 'Number.206': 0.0, 'Start.1': 0.0, 'Start.2': 0.0, 'Start.4': 0.0, 'Start.8': 0.0, 'Start.9': 0.0,
 'Start.11': 0.0, 'Start.15': 0.0, 'Start.17': 0.0, 'Start.18': 0.0, 'Start.20': 0.0, 'Start.22': 0.0, 'Start.2
6': 0.0, 'Start.27': 0.0, 'Start.31': 0.0, 'Start.35': 0.0, 'Start.36': 0.0, 'Start.37': 0.0, 'Start.42': 0.0,
 'Start.51': 0.0, 'Start.52': 0.0, 'Start.59': 0.0, 'Start.61': 0.0, 'Start.68': 0.0, 'Start.71': 0.0, 'Start.7
2': 0.0, 'Start.73': 0.0, 'Start.78': 0.0, 'Start.80': 0.0, 'Start.81': 0.0, 'Start.82': 0.0, 'Start.87': 0.0,
 'Start.91': 0.0, 'Start.93': 0.0, 'Start.96': 0.0, 'Start.97': 0.0, 'Start.100': 0.0, 'Start.102': 0.0, 'Star
t.105': 0.0, 'Start.112': 0.0, 'Start.113': 0.0, 'Start.114': 0.0, 'Start.118': 0.0, 'Start.120': 0.0, 'Start.1
21': 0.0, 'Start.125': 0.0, 'Start.127': 0.0, 'Start.128': 0.0, 'Start.130': 0.0, 'Start.131': 0.0, 'Start.13
6': 0.0, 'Start.139': 0.0, 'Start.140': 0.0, 'Start.143': 0.0, 'Start.148': 0.0, 'Start.151': 0.0, 'Start.157':
 0.0, 'Start.158': 0.0, 'Start.159': 0.0, 'Start.168': 0.0, 'Start.175': 0.0, 'Start.177': 0.0, 'Start.178': 0.
0, 'Start.195': 0.8124189135387114, 'Start.206': 0.0}, 'std_coefs_class_206': {'Intercept': -4.037349272241172,
 'Age.1': 0.0, 'Age.2': 0.0, 'Age.4': 0.0, 'Age.8': 0.0, 'Age.9': 0.0, 'Age.11': 0.0, 'Age.15': 0.0, 'Age.17':
 0.0, 'Age.18': 0.0, 'Age.20': 0.0, 'Age.22': 0.0, 'Age.26': 0.0, 'Age.27': 0.0, 'Age.31': 0.0, 'Age.35': 0.0,
 'Age.36': 0.0, 'Age.37': 0.0, 'Age.42': 0.0, 'Age.51': 0.0, 'Age.52': 0.0, 'Age.59': 0.0, 'Age.61': 0.0, 'Age.
68': 0.0, 'Age.71': 0.0, 'Age.72': 0.0, 'Age.73': 0.0, 'Age.78': 0.0, 'Age.80': 0.0, 'Age.81': 0.0, 'Age.82':
 0.0, 'Age.87': 0.0, 'Age.91': 0.0, 'Age.93': 0.0, 'Age.96': 0.0, 'Age.97': 0.0, 'Age.100': 0.0, 'Age.102': 0.
0, 'Age.105': 0.0, 'Age.112': 0.0, 'Age.113': 0.0, 'Age.114': 0.0, 'Age.118': 0.0, 'Age.120': 0.0, 'Age.121':
 0.0, 'Age.125': 0.0, 'Age.127': 0.0, 'Age.128': 0.0, 'Age.130': 0.0, 'Age.131': 0.0, 'Age.136': 0.0, 'Age.13
9': 0.0, 'Age.140': 0.0, 'Age.143': 0.0, 'Age.148': 0.0, 'Age.151': 0.0, 'Age.157': 0.0, 'Age.158': 0.0, 'Age.1
59': 0.0, 'Age.168': 0.0, 'Age.175': 0.0, 'Age.177': 0.0, 'Age.178': 0.0, 'Age.195': 0.0, 'Age.206': 0.81241912
68912065, 'Number.1': 0.0, 'Number.2': 0.0, 'Number.4': 0.0, 'Number.8': 0.0, 'Number.9': 0.0, 'Number.11': 0.
0, 'Number.15': 0.0, 'Number.17': 0.0, 'Number.18': 0.0, 'Number.20': 0.0, 'Number.22': 0.0, 'Number.26': 0.0,
 'Number.27': 0.0, 'Number.31': 0.0, 'Number.35': 0.0, 'Number.36': 0.0, 'Number.37': 0.0, 'Number.42': 0.0, 'N
umber.51': 0.0, 'Number.52': 0.0, 'Number.59': 0.0, 'Number.61': 0.0, 'Number.68': 0.0, 'Number.71': 0.0, 'Numb
er.72': 0.0, 'Number.73': 0.0, 'Number.78': 0.0, 'Number.80': 0.0, 'Number.81': 0.0, 'Number.82': 0.0, 'Number.
87': 0.0, 'Number.91': 0.0, 'Number.93': 0.0, 'Number.96': 0.0, 'Number.97': 0.0, 'Number.100': 0.0, 'Number.10
2': 0.0, 'Number.105': 0.0, 'Number.112': 0.0, 'Number.113': 0.0, 'Number.114': 0.0, 'Number.118': 0.0, 'Numbe
r.120': 0.0, 'Number.121': 0.0, 'Number.125': 0.0, 'Number.127': 0.0, 'Number.128': 0.0, 'Number.130': 0.0, 'Nu
mber.131': 0.0, 'Number.136': 0.0, 'Number.139': 0.0, 'Number.140': 0.0, 'Number.143': 0.0, 'Number.148': 0.0,
 'Number.151': 0.0, 'Number.157': 0.0, 'Number.158': 0.0, 'Number.159': 0.0, 'Number.168': 0.0, 'Number.175':
 0.0, 'Number.177': 0.0, 'Number.178': 0.0, 'Number.195': 0.0, 'Number.206': 0.8124191268912059, 'Start.1': 0.
0, 'Start.2': 0.0, 'Start.4': 0.0, 'Start.8': 0.0, 'Start.9': 0.0, 'Start.11': 0.0, 'Start.15': 0.0, 'Start.1
7': 0.0, 'Start.18': 0.0, 'Start.20': 0.0, 'Start.22': 0.0, 'Start.26': 0.0, 'Start.27': 0.0, 'Start.31': 0.0,
 'Start.35': 0.0, 'Start.36': 0.0, 'Start.37': 0.0, 'Start.42': 0.0, 'Start.51': 0.0, 'Start.52': 0.0, 'Start.5
9': 0.0, 'Start.61': 0.0, 'Start.68': 0.0, 'Start.71': 0.0, 'Start.72': 0.0, 'Start.73': 0.0, 'Start.78': 0.0,
 'Start.80': 0.0, 'Start.81': 0.0, 'Start.82': 0.0, 'Start.87': 0.0, 'Start.91': 0.0, 'Start.93': 0.0, 'Start.9
6': 0.0, 'Start.97': 0.0, 'Start.100': 0.0, 'Start.102': 0.0, 'Start.105': 0.0, 'Start.112': 0.0, 'Start.113':
 0.0, 'Start.114': 0.0, 'Start.118': 0.0, 'Start.120': 0.0, 'Start.121': 0.0, 'Start.125': 0.0, 'Start.127': 0.
0, 'Start.128': 0.0, 'Start.130': 0.0, 'Start.131': 0.0, 'Start.136': 0.0, 'Start.139': 0.0, 'Start.140': 0.0,
 'Start.143': 0.0, 'Start.148': 0.0, 'Start.151': 0.0, 'Start.157': 0.0, 'Start.158': 0.0, 'Start.159': 0.0, 'S
tart.168': 0.0, 'Start.175': 0.0, 'Start.177': 0.0, 'Start.178': 0.0, 'Start.195': 0.0, 'Start.206': 0.81241912
68912061}}

Coefficients: glm multinomial coefficients

	names	coefs_class_0	coefs_class_1	coefs_class_2	coefs_class_3	coefs_class_4	coefs_class_5	coefs_class_6	coefs_class_7	coefs_class_8
0	Intercept	-3.654539	-3.731125	-4.037075	-4.037140	-3.807430	-4.037260	-3.807528	-4.036455	-3.807548
1	Age.1	1.639074	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	Age.2	0.000000	1.453385	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	Age.4	0.000000	0.000000	0.812450	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	Age.8	0.000000	0.000000	0.000000	0.812448	0.000000	0.000000	0.000000	0.000000	0.000000
5	Age.9	0.000000	0.000000	0.000000	0.000000	1.275413	0.000000	0.000000	0.000000	0.000000
6	Age.11	0.000000	0.000000	0.000000	0.000000	0.000000	0.812446	0.000000	0.000000	0.000000
7	Age.15	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.275409	0.000000	0.000000
8	Age.17	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.813035	0.000000
9	Age.18	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.275408
10	Age.20	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
11	Age.22	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
12	Age.26	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
13	Age.27	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
14	Age.31	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
15	Age.35	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
16	Age.36	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
17	Age.37	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
18	Age.42	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
19	Age.51	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

20 rows × 129 columns

```
See the whole table with table.as_data_frame()
```

In [60]:

```
# define model
model = DecisionTreeClassifier()
# evaluate pipeline
# define pipeline
steps = [('over', SMOTE()), ('model', DecisionTreeClassifier())]
pipeline = Pipeline(steps=steps)
# evaluate pipeline
cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
scores = cross_val_score(pipeline, X, y, scoring='roc_auc', cv=cv, n_jobs=-1)
print('Mean ROC AUC: ', scores.mean())
```

```
Mean ROC AUC:  0.8123015873015874
```

Previously Descision tree accuracy was 0.63. So it is better