Exploratory Data Analysis

- In this notebook we will try to explore and understand the Haberman dataset.
- This dataset contains information about the patients who had undergone surgery for breast cancer

Objective:

Objective is to find out how can we classify the patient whether he can survive more than 5 years or not.

In [1]:
```
import pandas as pd
import numpy as np
import cufflinks as cf
from plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot
init_notebook_mode(connected = True) # to get the connection
cf.go_offline() #offline mode


df=pd.read_csv('go/EDA/haberman.csv',names=['age','op_year','axil_nodes','survival_status'])
```

In [2]:
```
print(df.head())
print(len(df))
print(len(df.columns))
print(df.columns.values)
print(df['survival_status'].unique())
print(df['survival_status'].value_counts())
```

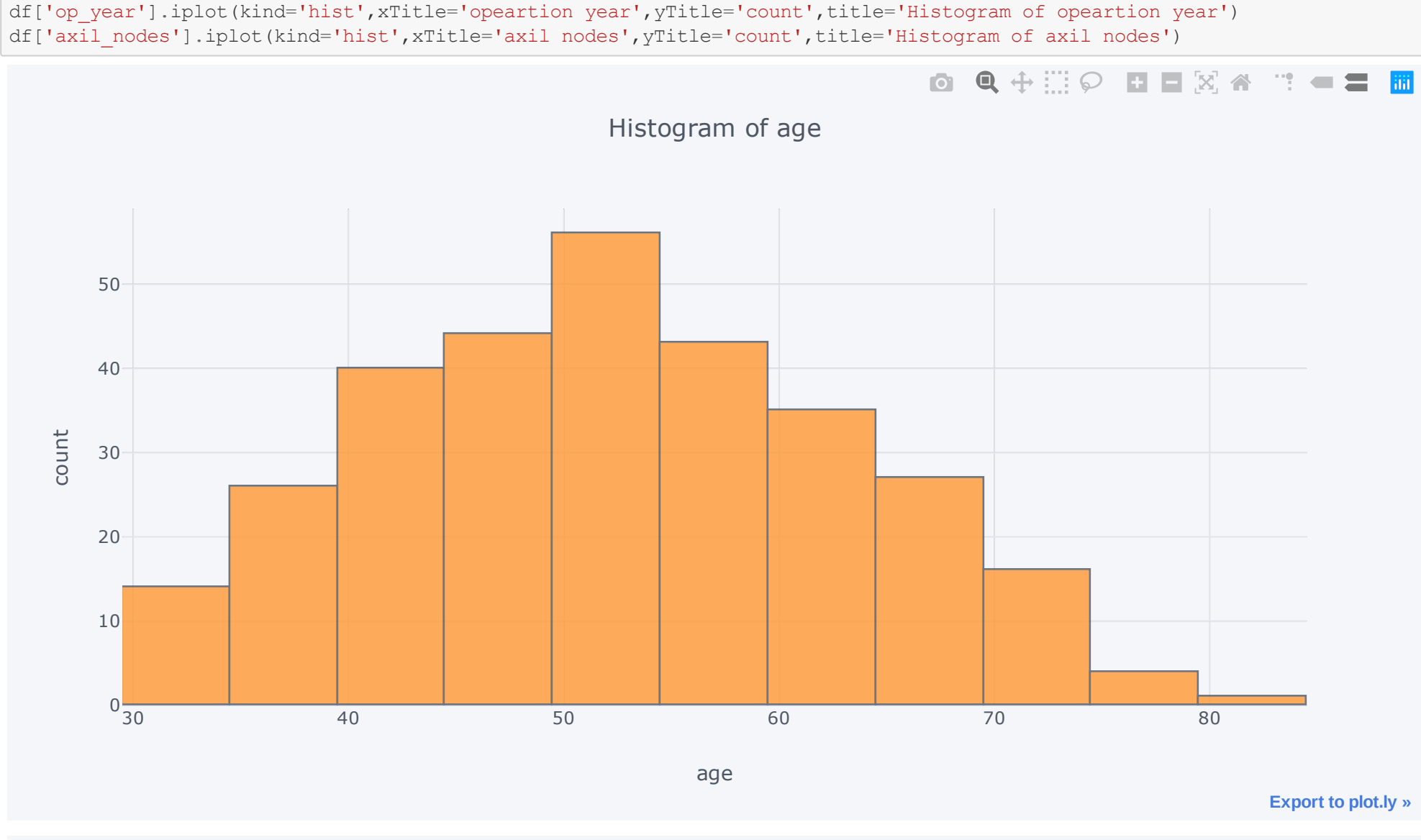
```
   age  op_year  axil_nodes  survival_status
0   30       64           1                1
1   30       62           3                1
2   30       65           0                1
3   31       59           2                1
4   31       65           4                1
306
4
['age' 'op_year' 'axil_nodes' 'survival_status']
[1 2]
1    225
2     81
Name: survival_status, dtype: int64
```

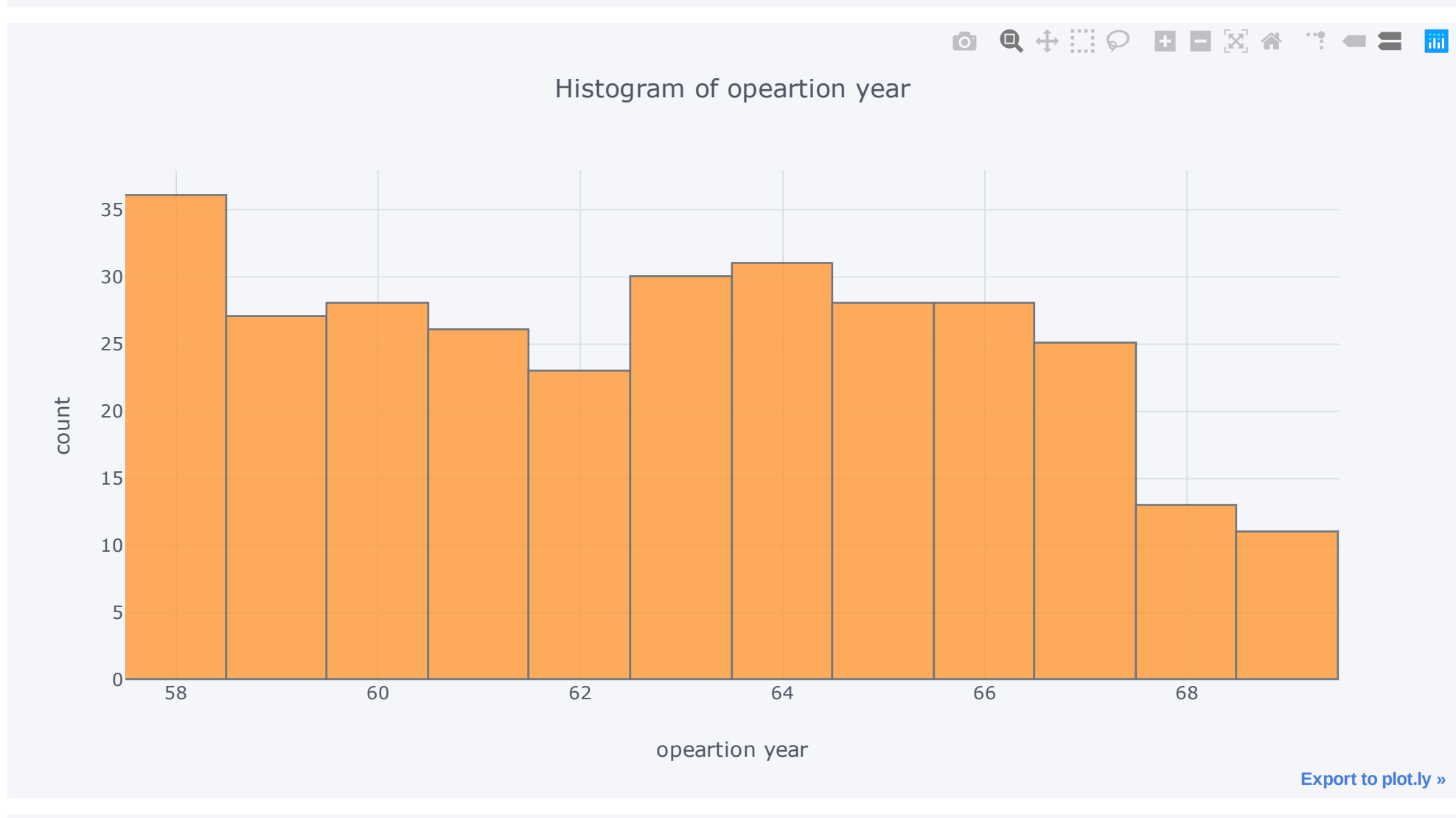
Conclusions:

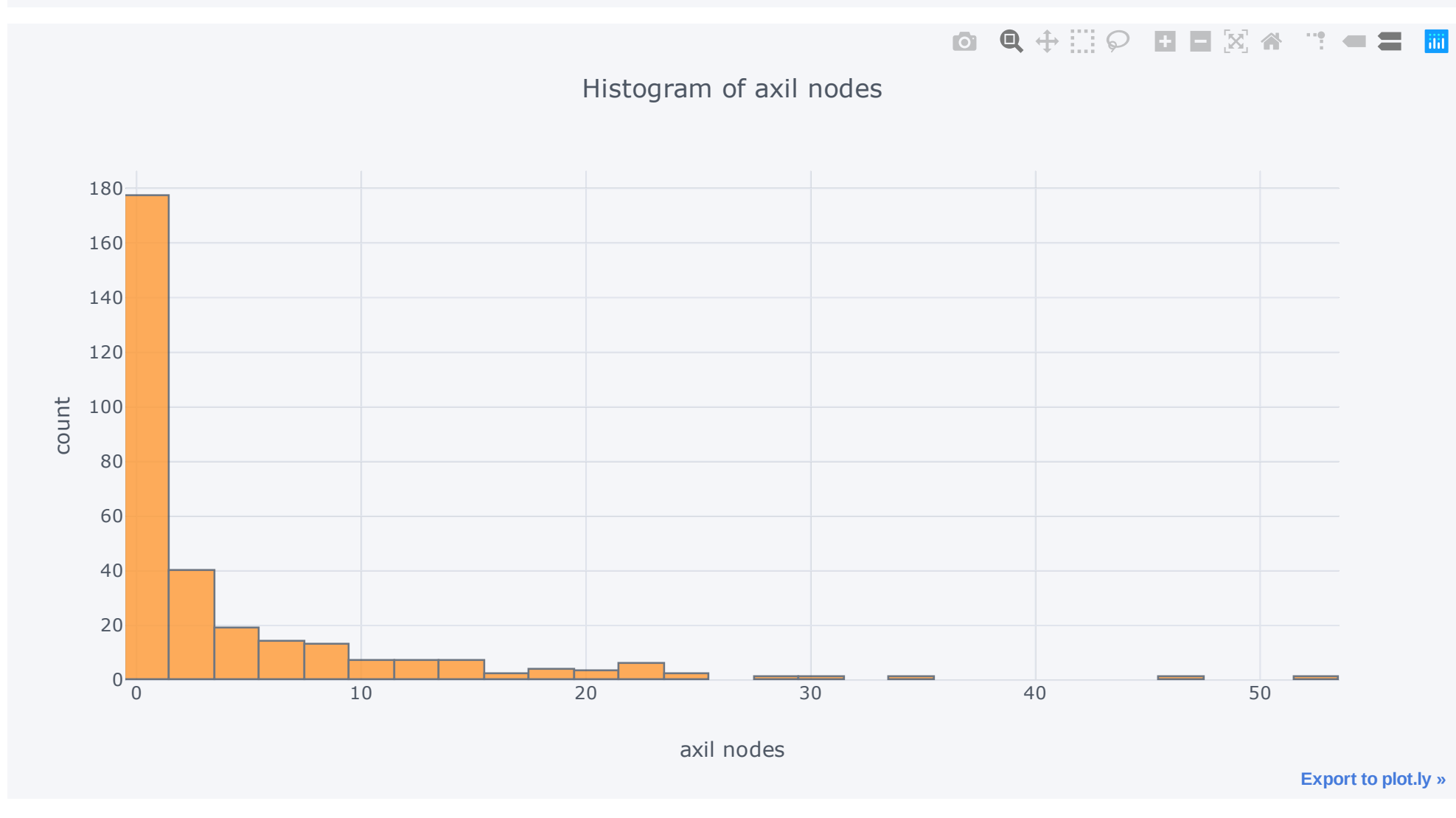
- The dataset contains 306 obervations.
- Total 4 columns are avilable in the dataset.Out of them we have age,op_year(operation year) and axil_nodes as features to predcit the survival_status.
- Number of classes to predict is: 2
- class '1' represents patient survived more than 5 years and '2' represents patient died within 5 years.
- we have 225 observations for class '2' and and 81 observations for class '1'.
- Dataset is slightly imbalanced.

Distributions of features

In [18]:
```
df['age'].iplot(kind='hist',xTitle='age',yTitle='count',title='Histogram of age')
df['op_year'].iplot(kind='hist',xTitle='opeartion year',yTitle='count',title='Histogram of opeartion year')
df['axil_nodes'].iplot(kind='hist',xTitle='axil nodes',yTitle='count',title='Histogram of axil nodes')
```







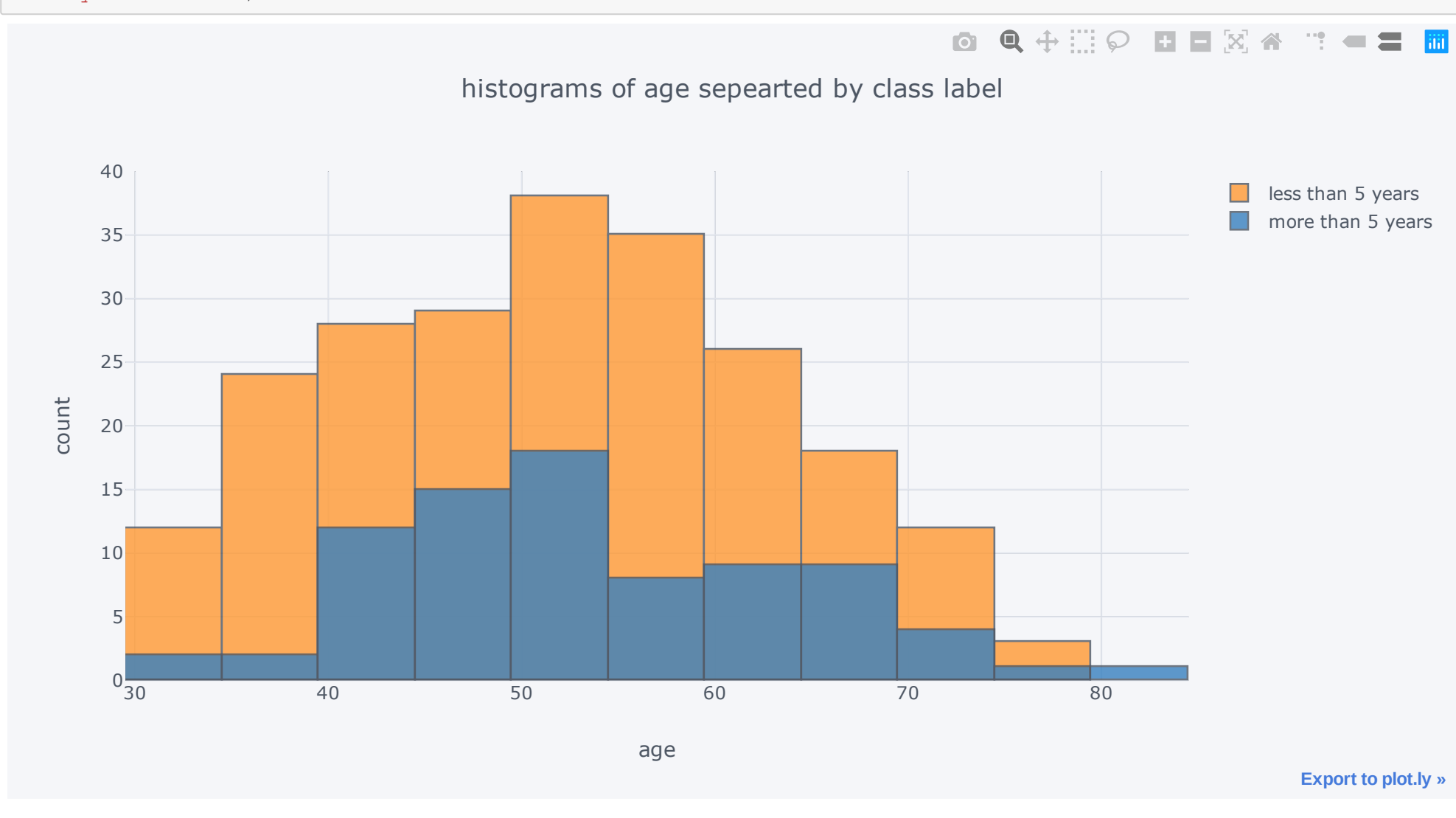
Conclusions:

- We can clearly observe age is following normal distribution.
- axil node feature is heavily skewed towards left(more values are on the left side of the mean).
- And operation year is not following any specific distribution.

Univariate Analysis

In [3]:
```
df['survival_status']=df['survival_status'].apply(lambda x:'more than 5 years' if x==2 else 'less than 5 years')
```

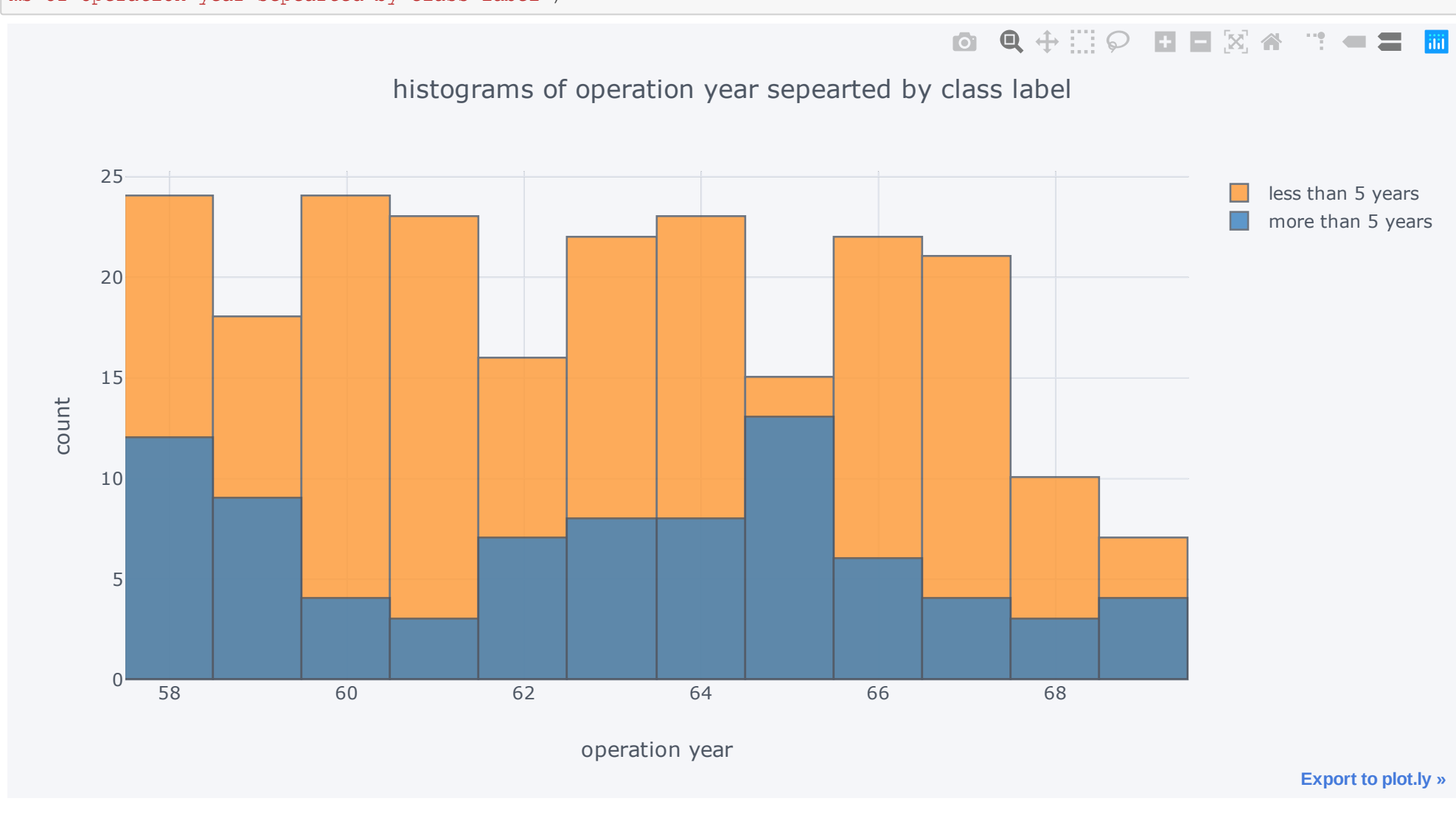
In [4]:
```
df.pivot(columns='survival_status',values='age').iplot(kind='hist',xTitle='age',yTitle='count',title='histograms of age sepea
rted by class label')
```



Conclusion:

- Age attribute alone can't distninguish between survival of patient.
- However we can see that lower age people having high chance og dying within 5 years.There are 40 people in the between 30 to 39 age.But only 4 of them are able to live more than 5 years.

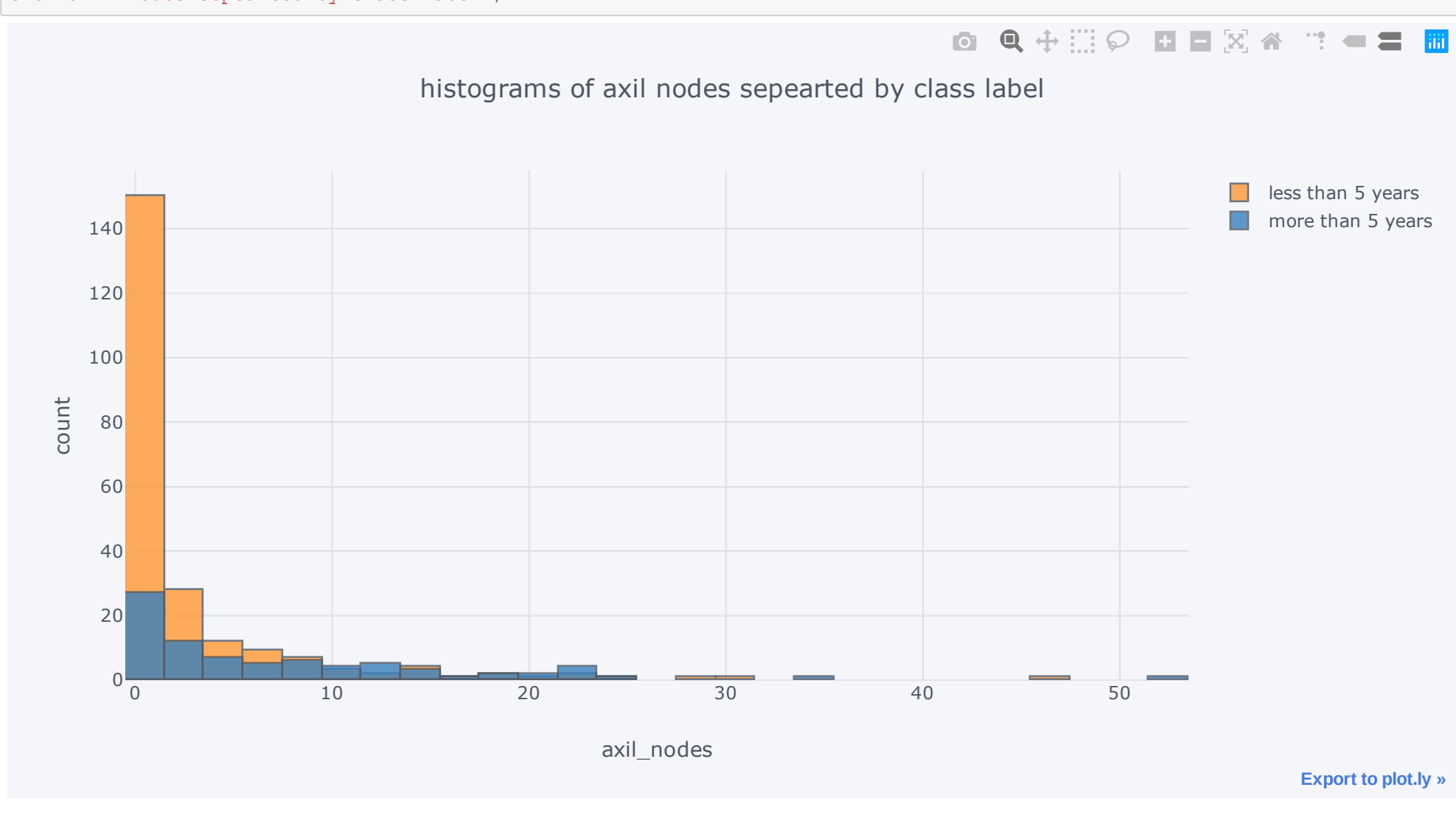
In [5]:
```
df.pivot(columns='survival_status',values='op_year').iplot(kind='hist',xTitle='operation year',yTitle='count',title='histogra
ms of operation year sepearted by class label')
```



Conclusions:

- Operation year alone attribute can't distniguish between two classes.
- Surprisingly the people who had surgery in 1965 have the highest chance of living more than 5 years.13 of 15 people are able to live more than 5 years.

In [6]:
```
df.pivot(columns='survival_status',values='axil_nodes').iplot(kind='hist',xTitle='axil_nodes',yTitle='count',title='histogram
s of axil nodes sepearted by class label')
```



conclusions:

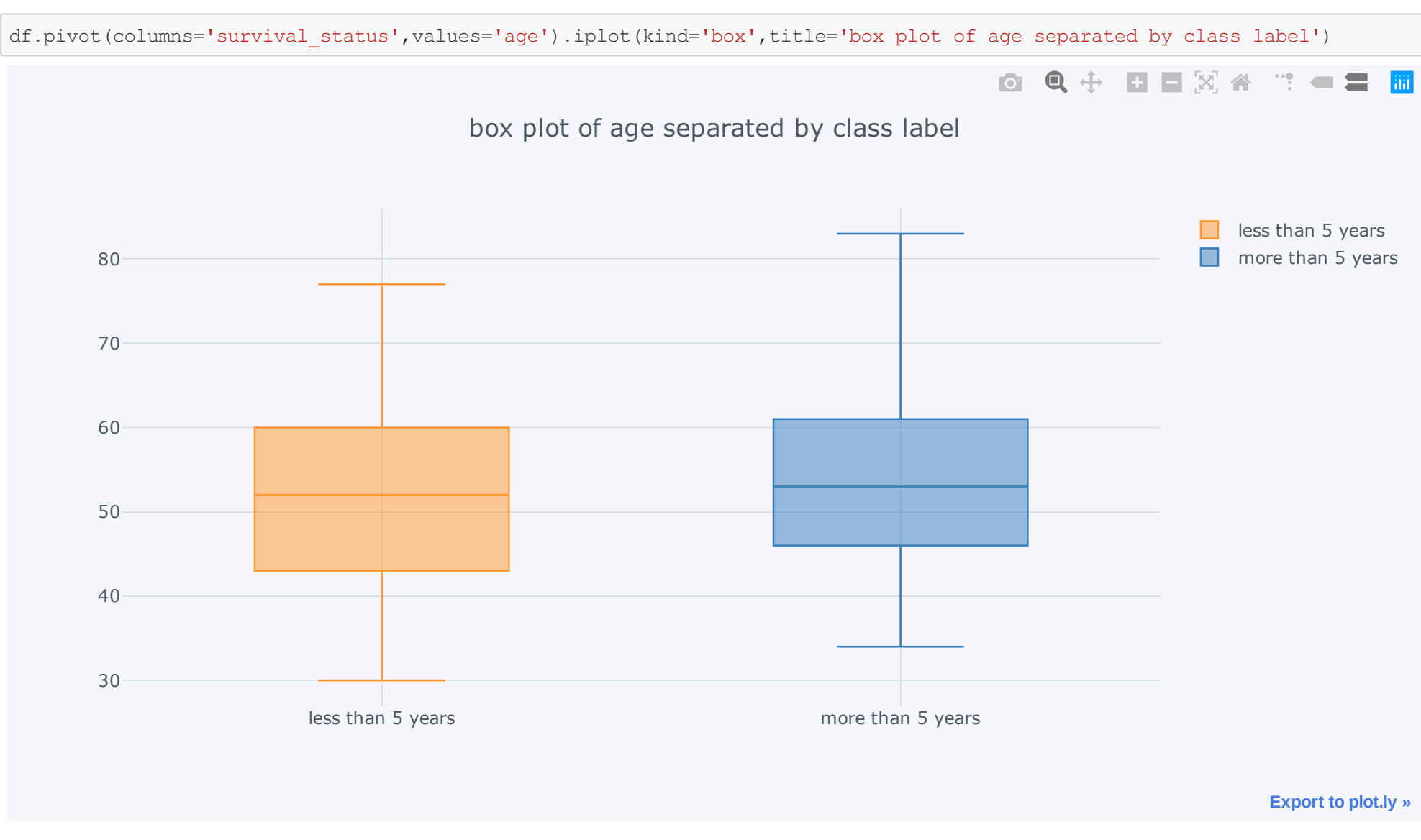
- axil_nodes also can't distniguish between two classes.
- There is higher chance of dying within 5 years if patients have lower number number of axil nodes.There are 40 people who are having 2/3 axil nodes out of them only 12 people are able to live more than 5 years.

In [7]:
```
#box plot
df.pivot(columns='survival_status',values='op_year').iplot(kind='box',title='box plot of operation year separated by class la
bel')
```

Conclusions:

- Box plot provides median,qunatitles very easily.
- qunatiles of operation_year for less than 5 year survival are:(58,60,63,66,69)
- qunatiles of operation_year for more than 5 year survival are:(58,59,63,65,69)

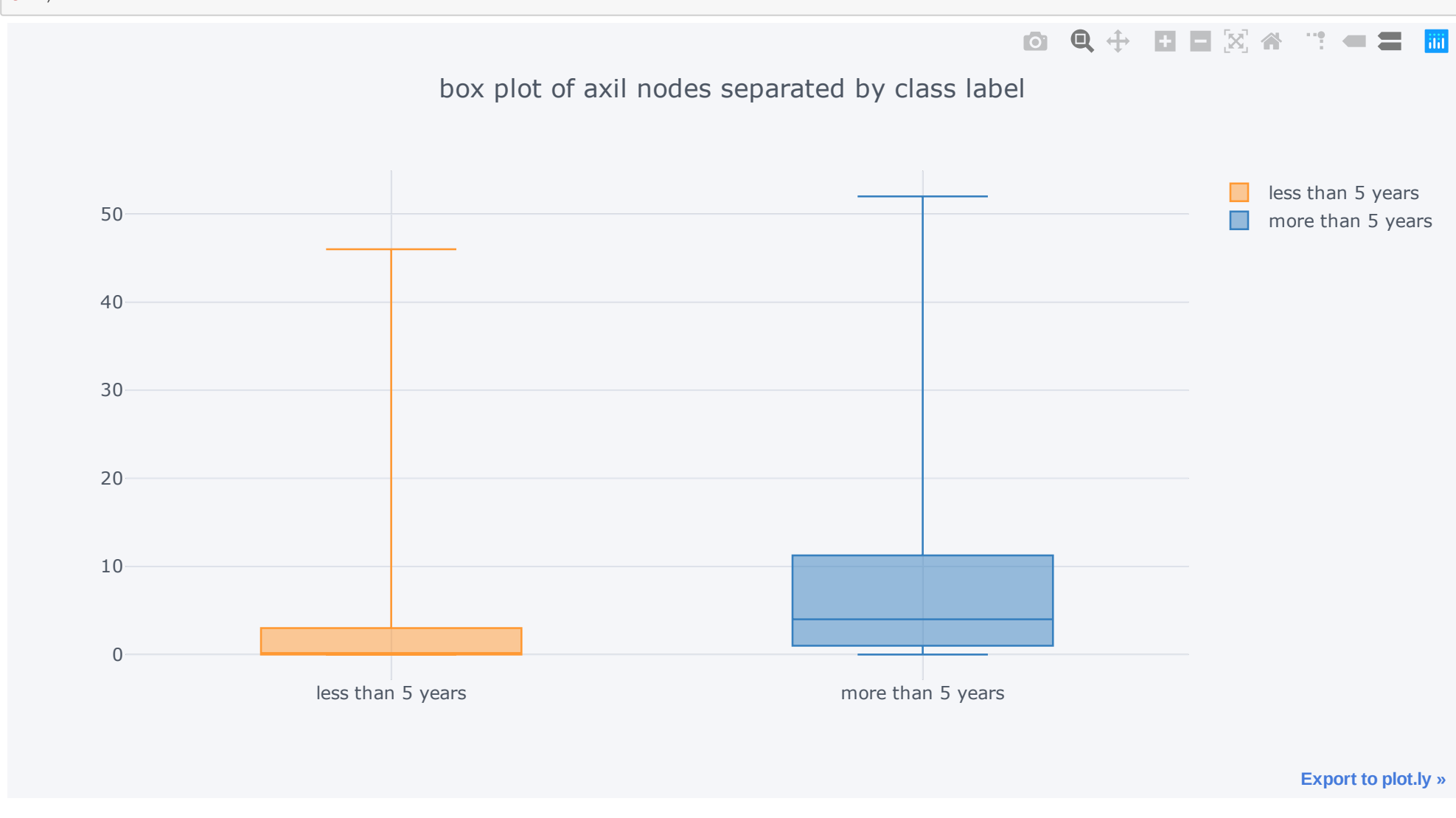
In [8]:
```
df.pivot(columns='survival_status',values='age').iplot(kind='box',title='box plot of age separated by class label')
```



Conclusions:

- qunatiles of age for less than 5 year survival are:(30,43,52,60,77)
- qunatiles of age for more than 5 year survival are:(34,46,53,61,83)

In [9]:
```
df.pivot(columns='survival_status',values='axil_nodes').iplot(kind='box',title='box plot of axil nodes separated by class lab
el')
```

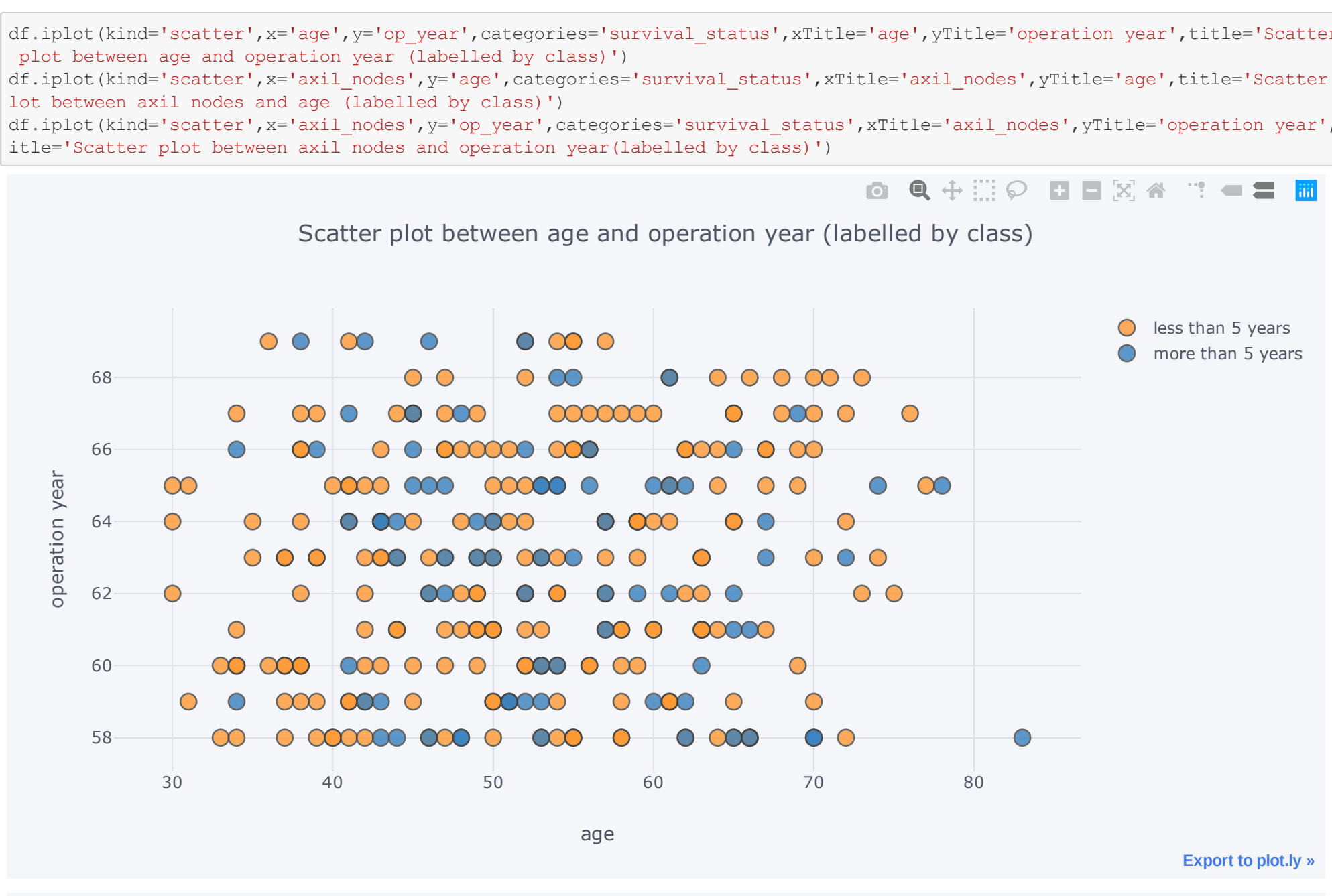


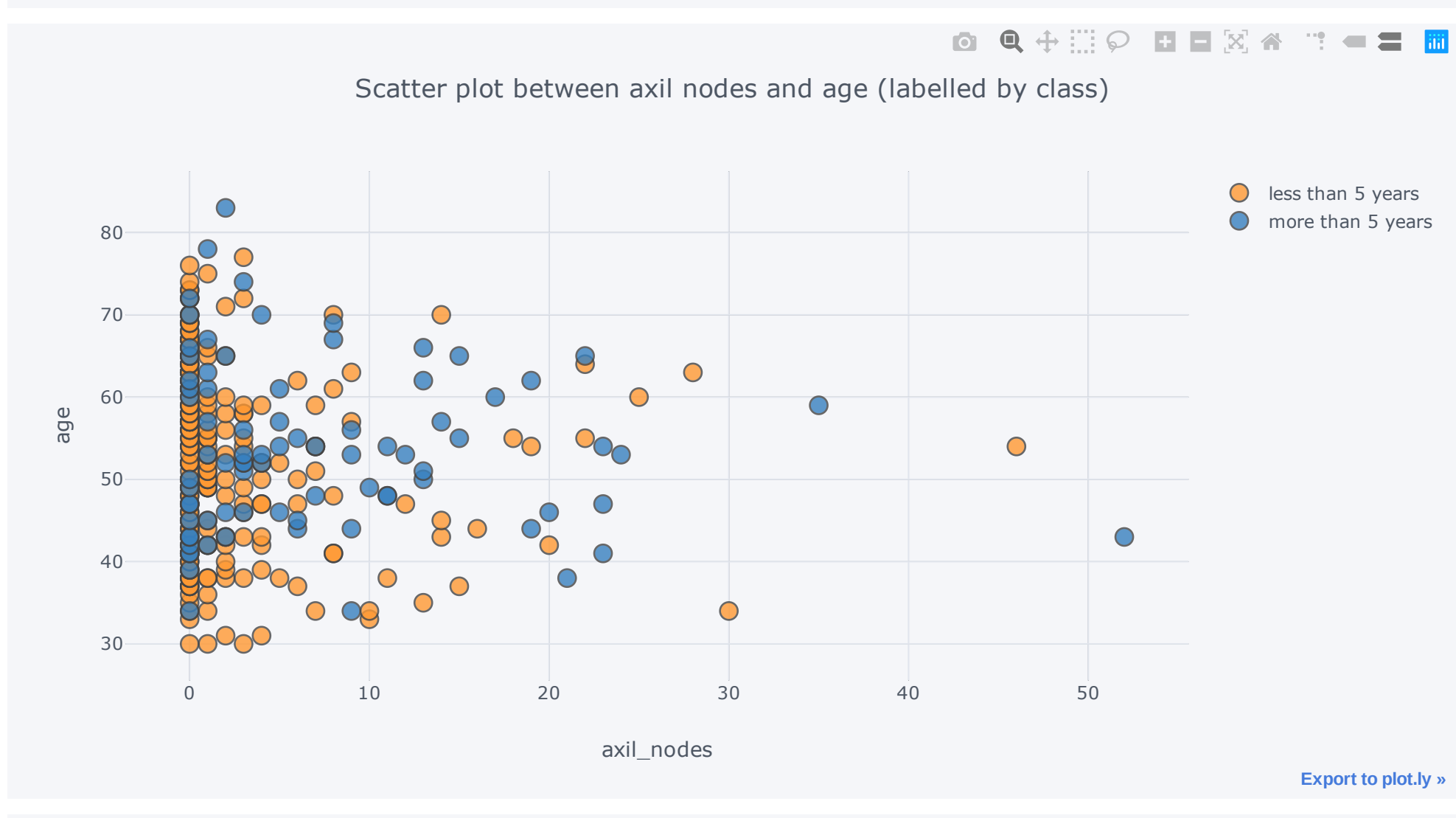
Conclusions:

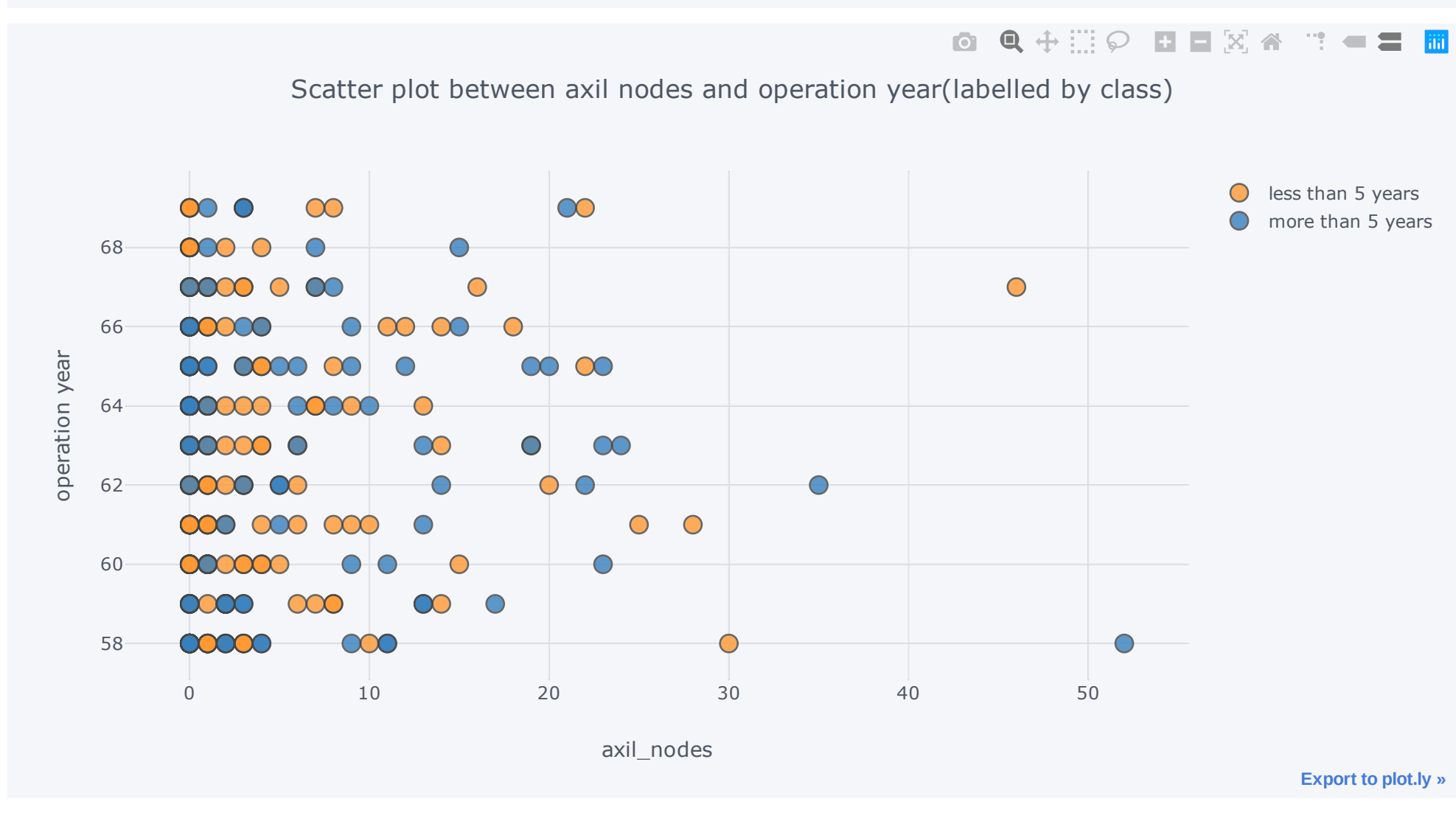
- 75 th percentile of axilnodes having less than 5 year survival is 3.
- median of axilnodes having less than 5 year survival is 0.It might refer that lower axil_nodes are causing to dying less than 5 years.
- 75 th percentile of axilnodes having more than 5 year survival is 11.25.
- maximum number of axil nodes is 52.

BiVariate Analysis

In [10]:
```
df.iplot(kind='scatter',x='age',y='op_year',categories='survival_status',xTitle='age',yTitle='operation year',title='Scatter
 plot between age and operation year (labelled by class)')
df.iplot(kind='scatter',x='axil_nodes',y='age',categories='survival_status',xTitle='axil_nodes',yTitle='age',title='Scatter p
lot between axil nodes and age (labelled by class)')
df.iplot(kind='scatter',x='axil_nodes',y='op_year',categories='survival_status',xTitle='axil_nodes',yTitle='operation year',t
itle='Scatter plot between axil nodes and operation year(labelled by class)')
```







Conclusions:

- Out of all attributes axil_node is able to classify better than remianing ones.
- No two attributes are able to completly classify the classes.

Multi Variate Analysis

In [11]:
```
df.iplot(kind='scatter3d',x='age',y='op_year',z='axil_nodes',categories='survival_status',xTitle='age',yTitle='operation Yea
r',zTitle='axil nodes',title='3D-sctatter plot with all features(labelled by class)')
#Please check jupyter notebook/HTML file for this one as it is not possible to view this one in pdf.
```

Conclusions:

- By observing the 3D-scatter plot we can find plane to classify the which will give better accuracy.
- Feature importance could be axil_nodes >op_year> age.