Week4 practical

Univariate and Bivariate Notes

Define Univariate

Univariate data refers to a dataset where each observation is associated

with only one variable.

This means it focuses on measuring or observing a single characteristic

or attribute for each individual in the dataset.

Example :

In [8]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
df=pd.read_csv('C:/Users/user/Downloads/kaggle/auto-mpg.csv')
df
```

Out[8]:

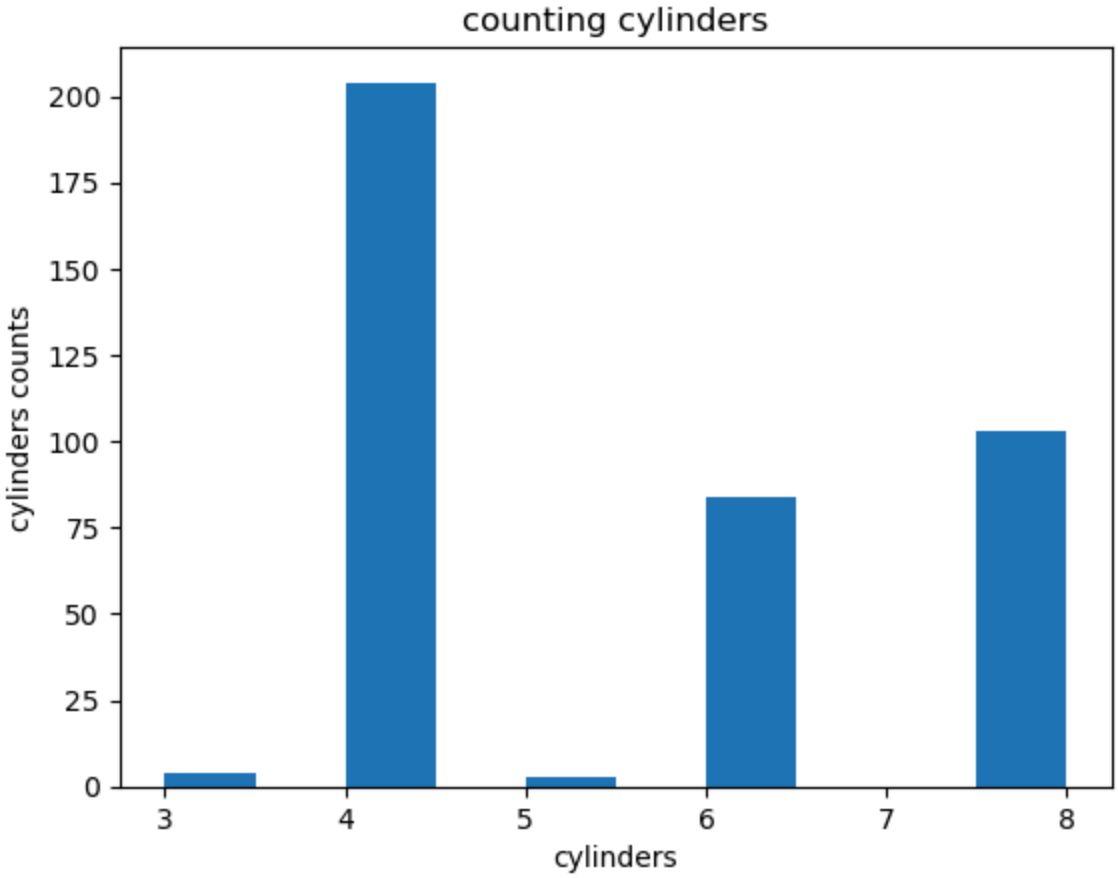
	mpg	cylinders	displacement	horsepower	weight	acceleration	model year
0	18.0	8	307.0	130	3504	12.0	70
1	15.0	8	350.0	165	3693	11.5	70
2	18.0	8	318.0	150	3436	11.0	70
3	16.0	8	304.0	150	3433	12.0	70
4	17.0	8	302.0	140	3449	10.5	70
...	...	...	...	...	...	...	...
393	27.0	4	140.0	86	2790	15.6	82
394	44.0	4	97.0	52	2130	24.6	82
395	32.0	4	135.0	84	2295	11.6	82
396	28.0	4	120.0	79	2625	18.6	82
397	31.0	4	119.0	82	2720	19.4	82

398 rows × 9 columns

Using univariate ploting Histogram

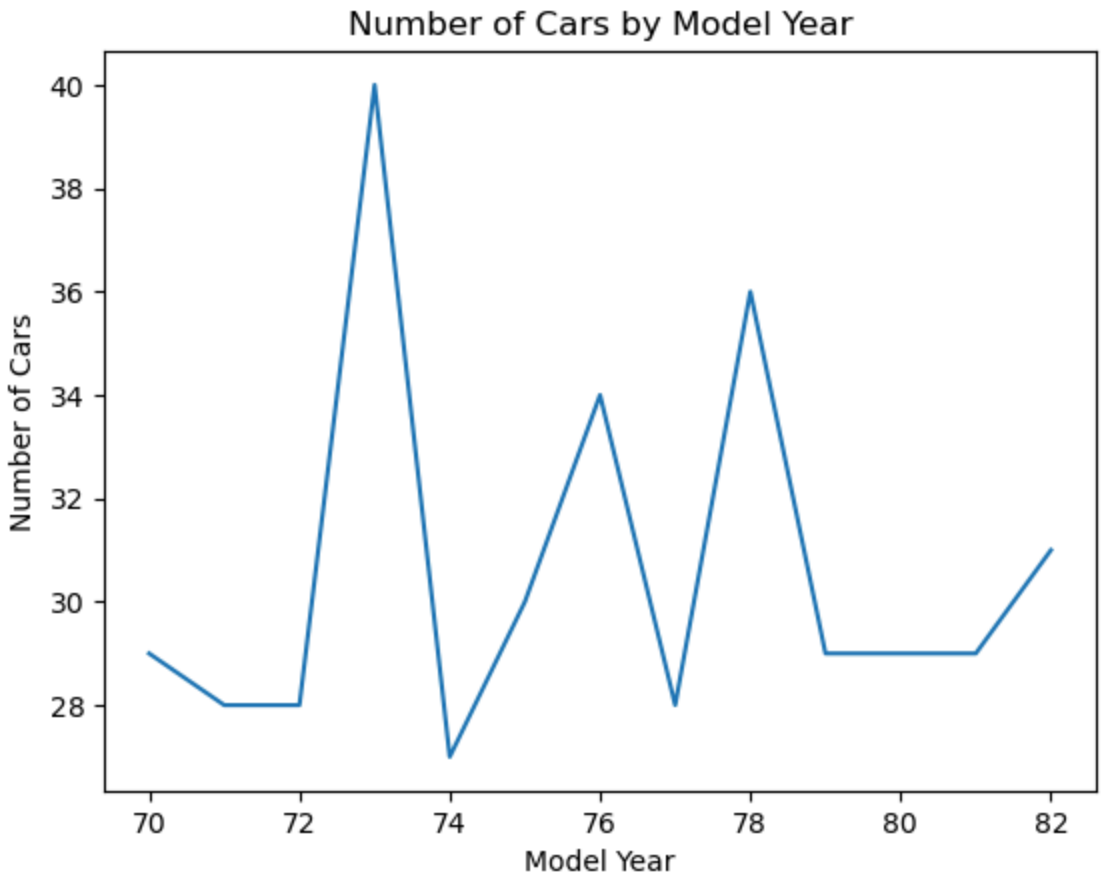
In [9]:
```
plt.hist(df['cylinders'])
plt.title('counting cylinders')
plt.xlabel('cylinders')
plt.ylabel('cylinders counts')
```

Out[9]:
```
Text(0, 0.5, 'cylinders counts')
```



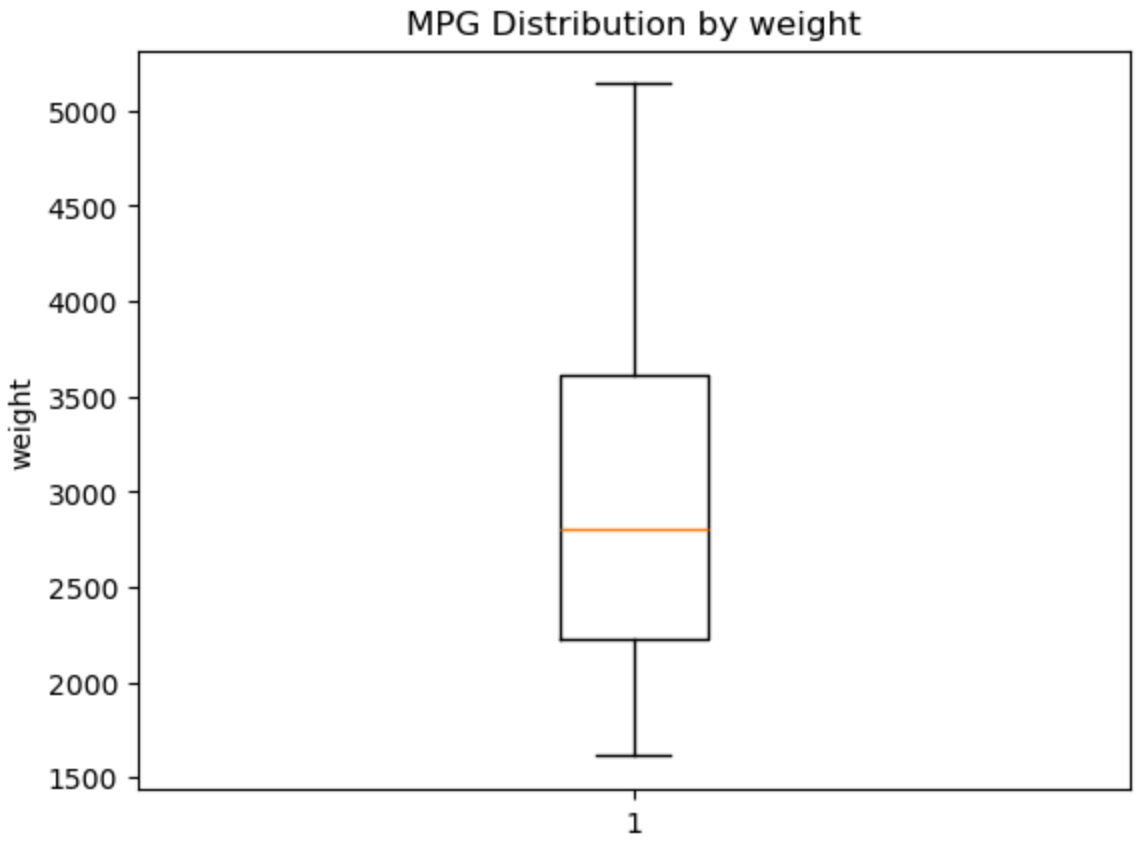
Using univariate ploting line graph

In [10]:
```
df['model year'].value_counts().sort_index().plot(kind='line')
plt.xlabel('Model Year')
plt.ylabel('Number of Cars')
plt.title('Number of Cars by Model Year')
plt.show()
```



Using univariate ploting boxplot

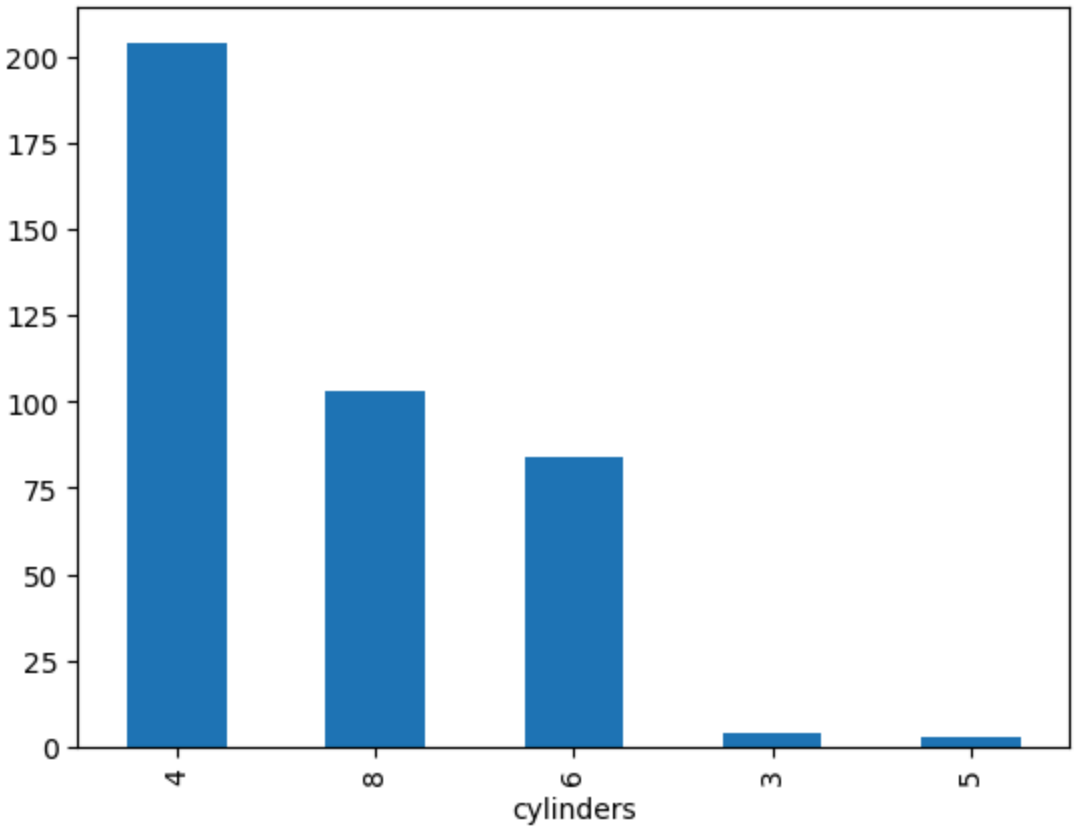
In [11]:
```
plt.boxplot(df['weight'])
plt.title('MPG Distribution by weight')
plt.ylabel('weight')
plt.show()
```



Using univariate ploting bar graph

In [12]:
```
df['cylinders'].value_counts().plot(kind='bar')
```

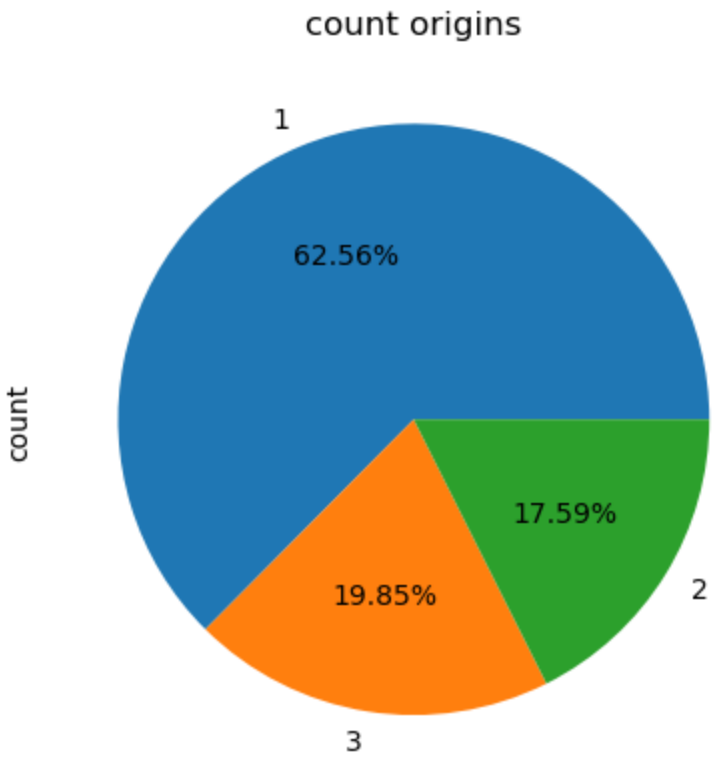
Out[12]:
```
<Axes: xlabel='cylinders'>
```



Using univariate ploting pie chart

In [13]:
```
df['origin'].value_counts().plot(kind='pie',autopct='%1.2f%%')
plt.title('count origins')
```

Out[13]:
```
Text(0.5, 1.0, 'count origins')
```



define Bivariate

Bivariate means the analysis of two variables.

Using bivariate analysis we can find how well the variables are correlated.

Bivariate analysis is of 3 types

1.Numerical variables

2.Categorical variables

3.Numerical & Categorical variable

Example:

Using titanic dataset

In [14]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
df=pd.read_csv('C:/Users/user/Downloads/kaggle/titanic (1).csv')
df
```

Out[14]:

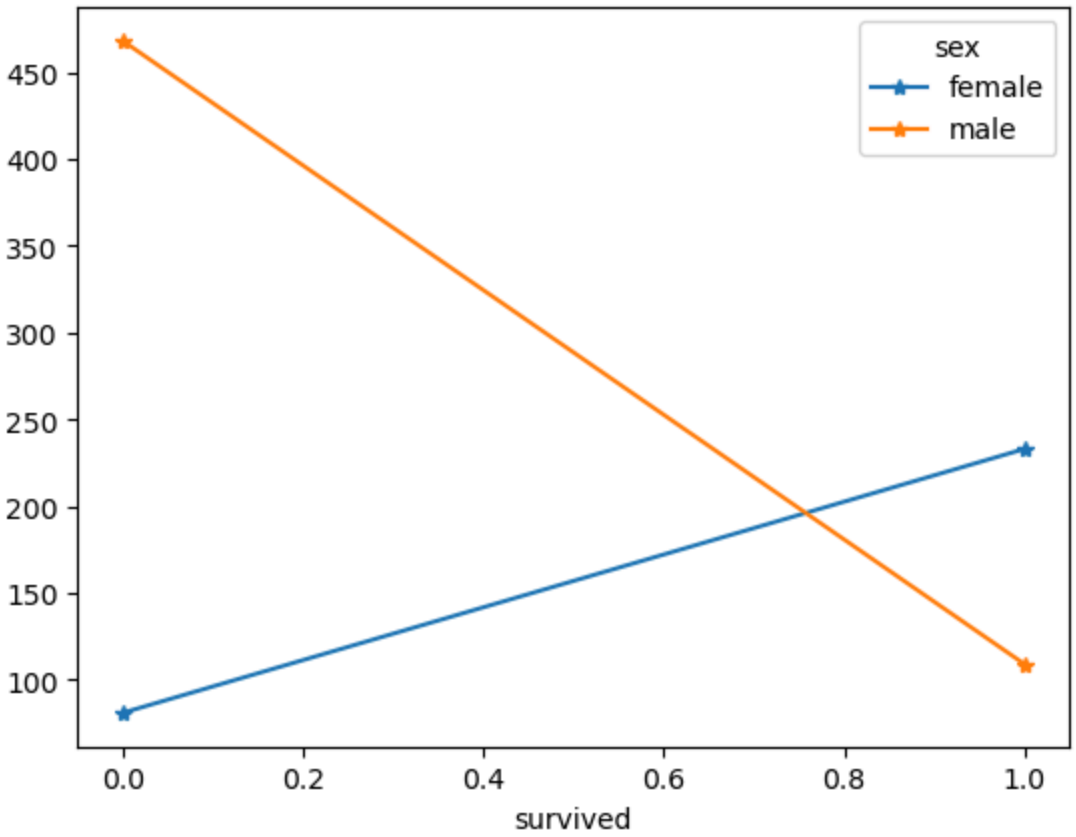
	survived	pclass	sex	age	sibsp	parch	fare	embarked	class
0	0	3	male	22.0	1	0	7.2500	S	Third
1	1	1	female	38.0	1	0	71.2833	C	First
2	1	3	female	26.0	0	0	7.9250	S	Third
3	1	1	female	35.0	1	0	53.1000	S	First
4	0	3	male	35.0	0	0	8.0500	S	Third
...	...	...	...	...	...	...	...	...	...
886	0	2	male	27.0	0	0	13.0000	S	Second
887	1	1	female	19.0	0	0	30.0000	S	First
888	0	3	female	NaN	1	2	23.4500	S	Third
889	1	1	male	26.0	0	0	30.0000	C	First
890	0	3	male	32.0	0	0	7.7500	Q	Third

891 rows × 15 columns

Using bivariate ploting line graph

In [15]:
```
a=df.groupby(['survived','sex']).size().unstack()
a.plot(kind='line',marker='*')
```

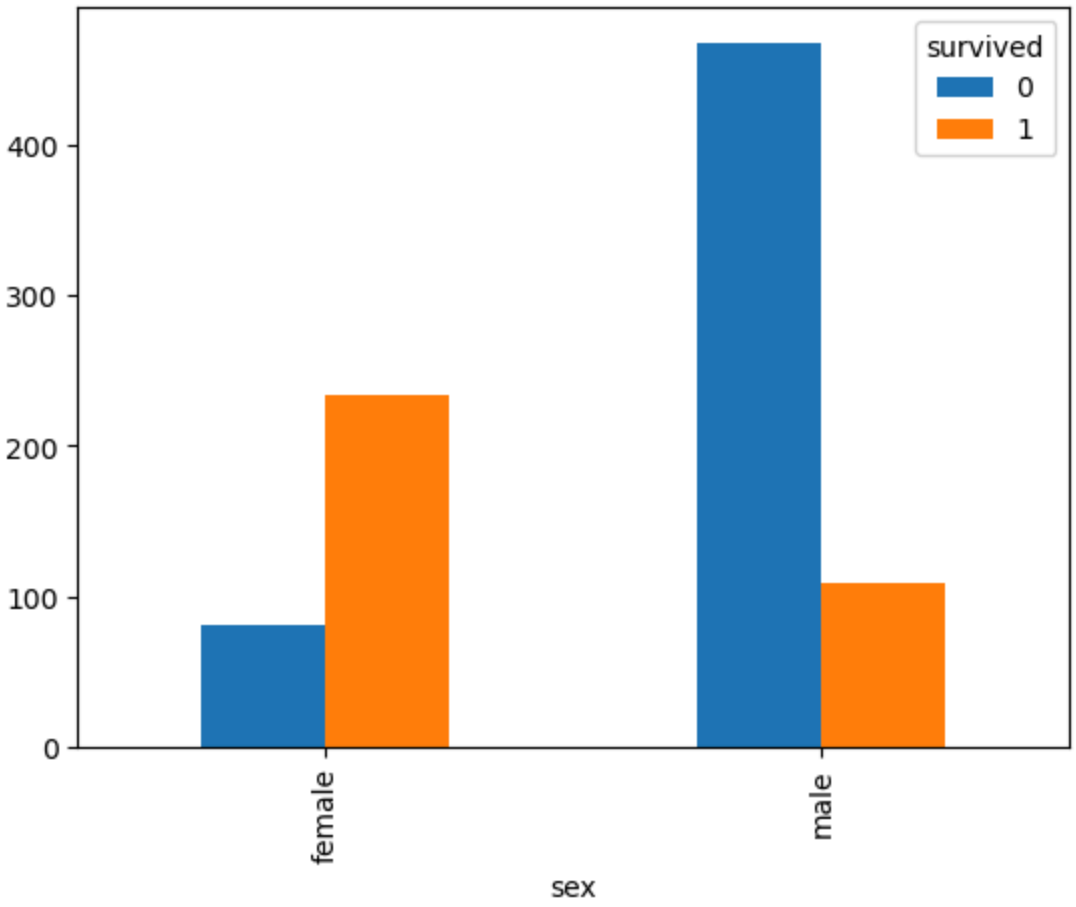
Out[15]:
```
<Axes: xlabel='survived'>
```



Using bivariate ploting bar plot

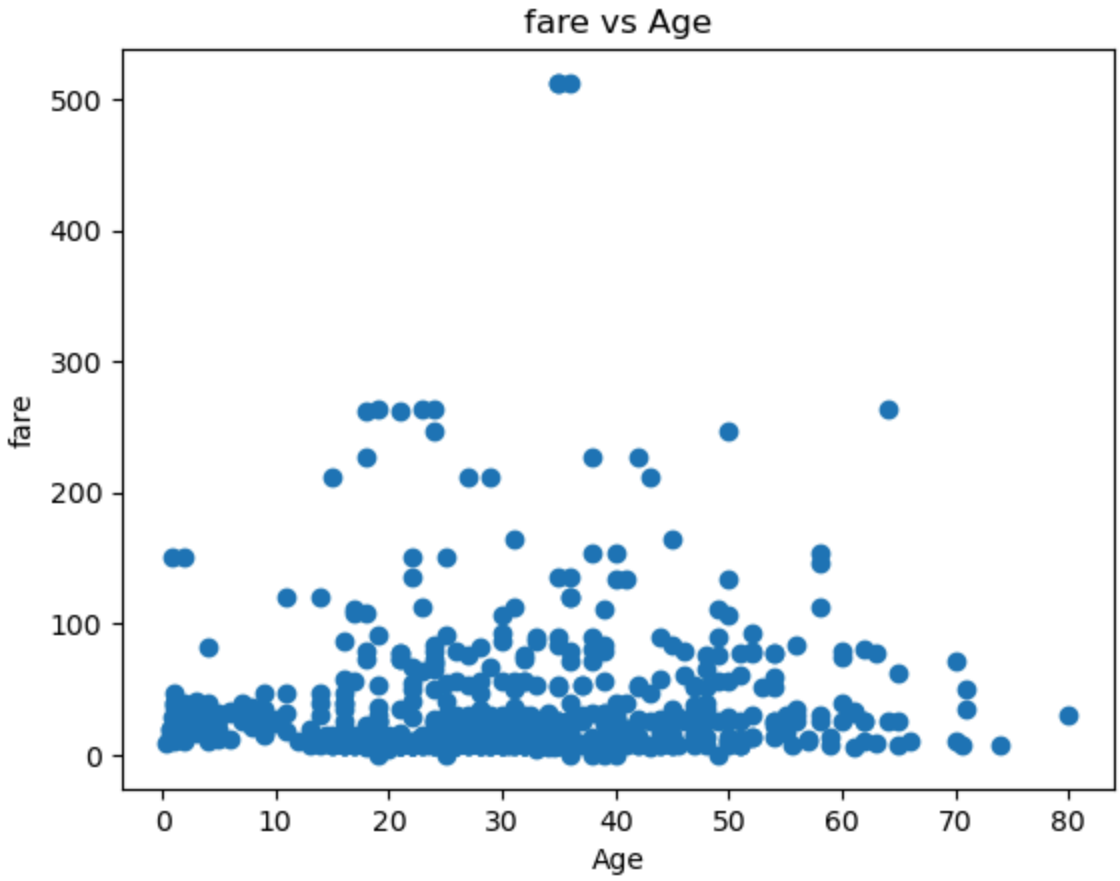
In [17]:
```
a=df.groupby(['sex','survived']).size().unstack()
a.plot(kind='bar')
```

Out[17]: <Axes: xlabel='sex'>



Using bivariate ploting scatter plot

In [7]:
```
plt.scatter(df['age'], df['fare'])
plt.xlabel('Age')
plt.ylabel('fare')
plt.title('fare vs Age')
plt.show()
```



In [ ]: