Project 3 "Passengers of the Titanic"

Pavel Olariu po3787

In [25]:
```
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
#allows multiple outputs for each code cell
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [29]:
```
titanic = sns.load_dataset('titanic')
titanic.head()
#loaded dataset and took a glimpse at columns and first few rows
```

Out[29]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True

In [18]:
```
titanic.info()
#gives generel info about dataset and variables
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
survived       891 non-null int64
pclass         891 non-null int64
sex            891 non-null object
age            714 non-null float64
sibsp          891 non-null int64
parch          891 non-null int64
fare           891 non-null float64
embarked       889 non-null object
class          891 non-null category
who            891 non-null object
adult_male     891 non-null bool
deck           203 non-null category
embark_town    889 non-null object
alive          891 non-null object
alone          891 non-null bool
dtypes: bool(2), category(2), float64(2), int64(4), object(5)
memory usage: 80.6+ KB
```

This dataset containts information about the RMS Titanic, a ship that sunk while crossing the Atlantic Ocean in 1912. The dataset contains 15 variables, which describe features of each individual passenger such as the amount they payed for fare (in USD$), their age in years, the passenger class they were a part of, their sex, if they survived the accident, and a few others that I will not focus on. There are 891 rows in this dataset, which each represents one individual passenger present on the Titanic at the time of sinkage.

In [20]:
```
titanic.describe()
#gives many descriptive statistics about numeric variables
```

Out[20]:

	survived	pclass	age	sibsp	parch	fare
count	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

The statistics table above shows some interesting generic statistics. The mean passenger class was 2.3, meaning that a vast majority was second, but mostly in the third class, which makes sense. The mean age was slighlty less than 30 years, which again makes sense for the time period, when life expectancy was much lower than today. The maximum and minimum ages are also interesting to see (80 years and less than 6 months, respectively). The fare is also an interesting variable to look at, as it has a large range, which some paying no fare, and the highest being more than 512 dollars, which is very high for the time period.

In [24]:
```
pd.crosstab(index=titanic['sex'], columns='count')
#gives the count of males and females on the boat
sextab = pd.crosstab(index=titanic['sex'], columns='count')
#find proportions after defining the totals and dividing by sum
sextab/sextab.sum()
titanic.loc[titanic['sex'] == "female", 'survived'].mean()
titanic.loc[titanic['sex'] == "male", 'survived'].mean()
```

Out[24]:

col_0	count
sex	
female	314
male	577

Out[24]:

col_0	count
sex	
female	0.352413
male	0.647587

Out[24]: 0.7420382165605095

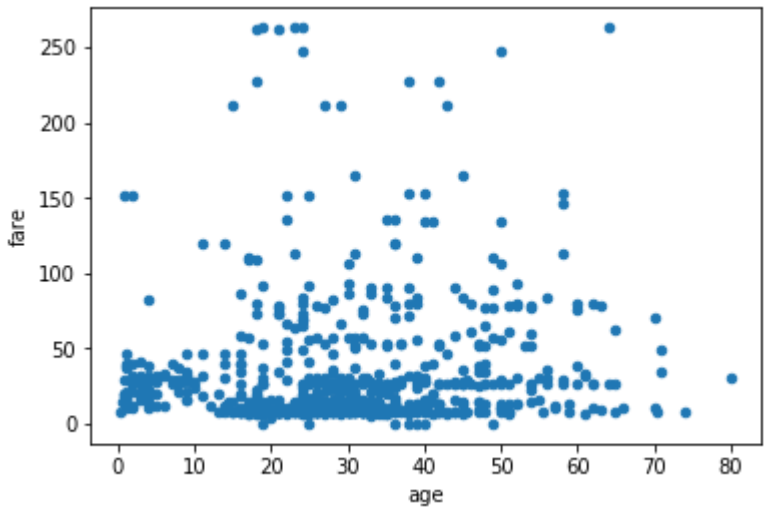
Out[24]: 0.18890814558058924

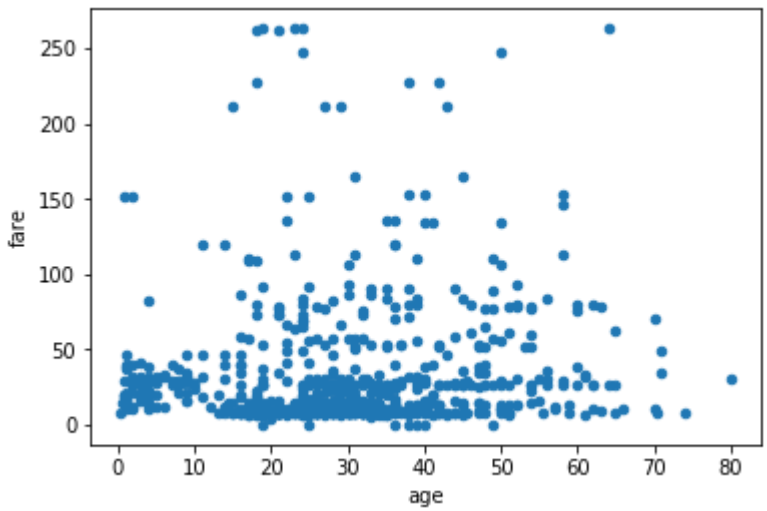
Looking at the sex variable, we can see that there are many more males on board than females (almost twice as many), which makes sense for the time period, as there was much sexism. But then looking at the survival rate of men versus women, we can see that the mean survival rate for women was about 74%, compared to men being about 19%. This is quite interesting as it gives some truth to the well known saying "Save the women and children first".

In [34]:
```
titanic.plot.scatter(x = 'age', y = 'fare')
#initial graph had two outliers that were above 500 dollars, so those values w
ere removed
titanic = titanic[titanic.fare != 512.329200]
titanic.plot.scatter(x = 'age', y = 'fare')
#created scatterplot comparing age and fare
```

Out[34]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa71f54ab38>

Out[34]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa725107128>

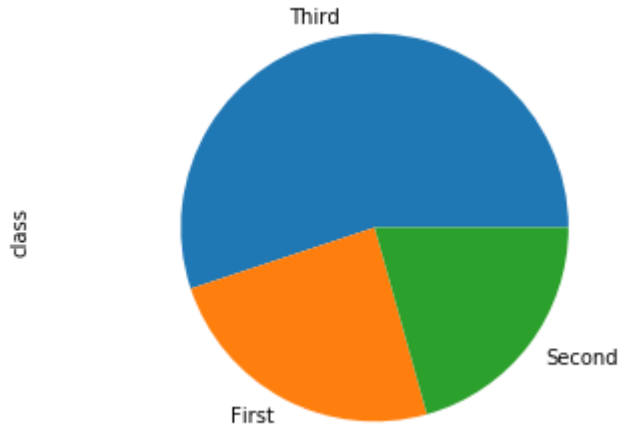




I was interested to see if there was any correlation or relation between the age of the passenger and the fare they were paying. As seen in the scatterplot above, there is not much correlation at all, and no pattern can clearly be seen. The biggest takeaway is that age did not have an affect on the fare rate, and the majority of fares were well below 50 dollars, which makes sense as most passengers were in the third class.

In [27]:
```
titanic['class'].value_counts() \
.plot(kind = "pie") \
.axis('equal') # equal aspect ratio
#pie chart created to look at pasenger classes
```

Out[27]:
```
(-1.110415878418142, 1.100496015606113, -1.134350102435046, 1.112420061514539)
```



In this pie chart we are able to see the distribution of passengers across classes. The vast majority of passengers were in the third class. There is about an equal amount of passengers in the first and second class, with first class having slightly more passengers than second class. This is interesting to see, because fare for crossing the ocean was not cheap, and this can be seen as a good representation of the economic and social classes in that time period. Most individuals were poor and of the lower class. The middle class was still fairly small, not the majority like it is today. And then first class was still quite large, as they could afford to make such journeys at leisure.